Revista Argentina de Clínica Psicológica
2026, Vol. XXXV, N°1, 87-104
DOI: 10.69980/03276716.2026.08
Received: 18 June 2026,
Acceptance: 03 August 2026

A Repairable M/M/1 Retrial Queue with Setup Times and Negative Arrivals: Conceptual Applications to Neural Information Processing and Cognitive Systems

P. Sukania[1], S. Pavai Madheswari[2], N. Meenatchi[3],
P. Suganthi[4]

Abstract

This study investigates a repairable M/M/1 retrial queue subjected to server setup times and the destabilizing effect of negative arrivals. The server is deactivated during idle periods to save energy. The server is reactivated once the primary customer arrives; however, it is not served immediately. The customer later joins a retrial orbit to attempt the service again. The analysis further integrates server breakdowns during active service periods, necessitating random repair intervals. A steady-state analysis is conducted using the probability-generating function method to derive the performance metrics, including the mean orbit length. Numerical experiments are conducted to assess the influence of various system parameters and to develop a cost-optimization model that identifies efficient operational policies, while considering energy consumption, retrial handling, and maintenance and repair costs. The developed framework is particularly applicable to systems where energy conservation and operational disruptions are critical, such as cloud computing infrastructures, telecommunication networks, and neural information-processing systems.

Key words: *M/M/1 Retrial Queue, Server Setup Times, Negative Arrivals, Server Breakdown and Repair, Idle–Time Server Shutdown, Probability-Generating Function, Cost Optimization, Cognitive Processing, and Neural Information Systems.*

INTRODUCTION

In real-world service systems, when a server is busy, it is regular for customers to wait, when the server is busy and try again later when the server is not busy. Rather than leaving altogether, retrial queue models are intended as they reattempt to receive the service after a certain amount of time. At the same time, there can be unusual arrivals, namely negative customers, which disturb the system by removing customers in service or even causing the server to stop working. The service disruptions make the system more chaotic and difficult to manage. To understand how the system performs, especially when both retrials and disruptions occur, it’s important to use models that consider all these components. This enables a better decision-making, which is led by a clear understanding of the system. These models simulate how people behave when immediate service is not available. Retrial queue models capture the performance and operational impact of repeated service attempts and system failures.

Queueing theory is a cornerstone for modelling and optimizing processes in diverse hybrid systems, from telecommunications and computing networks to automated manufacturing and service operations. In retrial queues, customers who find a server unavailable may not simply cease to exit but instead join an orbit to reattempt service after a random time. The intractability and feasibility of these models are further enhanced by other features such as server unreliability, strategic customer behaviour, and energy-saving strategies, optimizing the balance of performance, cost, and quality of service in contemporary infrastructures.

A key area of research focuses on integrating server setup times with retrial to model energy-efficient systems. Analytical developments in retrial queueing systems with reliability factors are illustrated by Nazarov et al. [21], a finite-source M/M/1 retrial queue with collisions, server breakdowns, and repairs are studied. The analysis provided rigorous characterization of system behavior in the heavy-traffic regime. This was subsequently extended to

[1]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: psukanianandakumar@gmail.com
[2]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: pavai.arunachalam@gmail.com
[3]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: nmi.sh@rmkec.ac.in
[4]Department of Mathematics, R.M.K. Engineering College, Tiruvallur, Tamil Nadu, India. E-mail: suganthirajkannu@gmail.com

provide realistic representations. Stochastic decomposition law was established by Madheswari et al. [15] investigated an M/G/1 retrial queue with a second optional service and customer balking under Bernoulli vacation schedules. Their study identified that vacation policies and service mechanisms play a crucial role in determining system stability and waiting-time distributions.

The latest studies have combined setup times with N-policies, where the server is activated only when a maximum number of customers arrive. Zhou et al. [35][36] investigated equilibrium joining strategies in retrial queues with setup times and N-policies, established the impact of setup delays on individual and socially optimal decisions. Economic control aspects were further explored by Zhang [33], analyzed optimal pricing in an unreliable retrial queue under two retrials. Multi-server retrial queues with delayed feedback, was examined by Melikov et al. [18] where an interrupted customer may either leave the system or join the orbit. Their analysis demonstrated that feedback delays play a crucial role in determining queue length and system availability measures. Reliability-oriented retrial queueing models were further extended to working vacations and multi-class customers by Muthusamy et al. [20], underscoring the interplay among reliability, vacation policies, and optimization objectives.

A combination of N-policy with multiple vacations was explored by Wang et al. [28] and derived equilibrium balking strategies under different levels. Their analysis identified the conditions under which individual rationality diverges from social optimal outcomes. Xu et al. [31] further combined pricing with Bernoulli preventive maintenance in an unreliable M/G/1 retrial queue, revealing the effectiveness of maintenance-aware pricing strategies. The insight of unreliability extends to systems with vacation policies, is similar as in the work of Shanmugam and Saravanarajan [23] and Bagyam et al. [2], who formalized queues with working vacations and repeated vacations. The reciprocal relationship between vacation policies and system performance represents inherent value in research.

Queueing theory investigated on sustainability, reliability, and customer response in single and multi-server systems in recent times. Jayamani et al. [10] proposed an N-policy M/M/1 queueing model to analyze energy-saving mechanisms in communication networks, demonstrating that adaptive control policies can considerably reduce power consumption while system performance is retained. Parallel to energy-aware modelling, the study of unreliable servers and repairable queueing systems has become an important topic in research. Server breakdowns are unavoidable in real-world service systems. Lv et al. [14] analysed an M/M/1 repairable queueing system with two types of server breakdowns and negative customers, revealing complex interactions between failure mechanisms and arrival processes.

Their work demonstrated the destabilizing effect of negative arrivals, was later extended to retrial settings with setup times by Tian et al. [26], yielding insights into comprehensive strategic analysis with breakdowns and repairs. Dhibar and Jain [8] adopted metaheuristics to analyze strategic behavior in retrial queues with breakdowns, vacations, and Bernoulli feedback, showing the effectiveness of heuristic optimization in complex systems. Parallel to this, statistical inference in queueing models has gained attention. Das et al. [7] and Bura and Sharma [4][5] examined the maximum likelihood and Bayesian estimation techniques for an M/M/1 queue with balking.

Discrete-time and non-Markovian extensions were explored by Wu and Lan [30], who analyzed repairable discrete-time queues with negative customers, disasters, balking, and working vacations. Energy-harvesting and cognitive radio applications were modeled by Sun et al. [25], who proposed a two-class retrial queue capturing strategic user behavior in green communication networks. Strategic priority and joining rules were examined by Wang et al. [29], highlighting the role of retrials in priority purchasing decisions.

Younes et al. [32] analysed the performance and reliability assessment of multi-service call admission control schemes in LTE networks. Hablal et al. [9], the latter providing stochastic bounds for priority retrial queues. Jeganathan et al. [11], analysed queueing-inventory systems with multiple service facilities. Ma et al. [16] and Guo et al. [17] proposed M/M/c queueing models with server vacation policies to improve virtual machine scheduling and energy efficiency in cloud computing. Bouchentouf et al. [3] designed a model for automated systems with dynamic vacations, waiting servers, customer impatience, and failures. Recent works emphasize advanced estimation and optimization. Singh et al. [24] extended the classical and Bayesian inference to single-server Markovian queues with arrivals during service times. Ke et al. [13] analyzed a feedback retrial queue with balking and server breakdowns, to determine optimal system parameters by developing a cost model. Cost optimization under N-policy and working breakdowns was analyzed by Vijayashree and Pavithra [27]. Intelligent and hybrid methods were proposed by Mehta and Jain [19], who applied ANFIS and metaheuristics to double-orbit retrial queues.

Kalaiselvi and Saravana Rajan [12] analyzed a non-Markovian single-server retrial queue with recurrent customers, balking, and extended Bernoulli vacations. Zhang and Wang [34] studied customer-intensive services with rational customers and retrials. Atencia et al. [1] developed a general non-Markovian retrial queueing framework. Fu-Min et al. [6] explored an unreliable-server retrial queue with customer feedback and impatience. Seenivasan et al. [22] used the matrix-geometric technique with catastrophes for a heterogeneous M/M/2 system.

The concurrence of retrial queueing theory is characterized by the integration of multiple realistic features such as setup times, server breakdowns, vacations, and customer feedback, linked with advanced economic and critical analyses.

On account of server failure, specific conditions where the server operates without clarity —it is essential to analyse customer behaviour and system performance in uncertain operating conditions. This paper considers a repairable M/M/1 retrial queue with setup times and disruptive negative arrivals. The server is shut down during inactive period and is setup on the arrival of a customer, who subsequently joins a retrial orbit, to conserve energy. Server breakdowns, triggered by negative customers, require random repair times. To obtain steady-state system properties and performance measures, including orbit queue length, technique of probability generating function is implemented. The precision of these measures is investigated numerically. A cost model is built to determine optimal system parameters for resource-efficient operation.

The System is particularly applicable to wireless communication networks, where the customers are data packets, and the servers correspond to the transmission nodes. When a node is occupied or inactive, the packets are queued in a retrial queue and retried based on the availability of the network. During active transmission, failures occur from issues such as signal fading or channel disruption. However, the inactive periods are used for energy saving due to the high energy demands of operating nodes. When new packets arrive, reactivation occurs, aligning them with energy-aware service strategies in real-world systems. Along with the signal delays in ATM networks, introduce setup periods comparable to the queueing delays considered in this study. These features emphasize the practical value of the model in capturing key characteristics in energy-efficient, fault-prone communication environments.

This paper provides a stochastic framework that can be used to represent certain aspects of cognitive and neural information processing, where retrials capture repeated memory access, negative arrivals represent interference and forgetting, and server breakdowns and repairs model cognitive fatigue and recovery processes.

NOTATION AND MODEL FORMULATION

Notations

λ	-	Arrival rate when system is idle.
λ_1	-	Arrival rate during busy time
λ_2	-	Arrival rate during setup time
λ_3	-	Arrival rate during Repair state
μ	-	Service rate
θ	-	Retrial rate
d	-	Probability for downtime
λ^-	-	Negative arrival during busy state
η	-	Repair rate
ξ	-	Breakdown rate
α	-	Setup rate
$\wp_o(z)$	-	Steady-state probabilities of the server being in an idle state
$\wp_1(z)$	-	Steady-state probabilities of the server being in a busy state
$\varpi\wp_2(z)$	-	Steady-state probability associated with the server's setup state
$\wp_3(z)$	-	Steady-state probability for the server's repair state
$\wp_o'(1)$	-	Mean number of customers in the orbit during an idle period
$\wp_1'(1)$	-	Mean number of customers in the orbit during a busy period
$\wp_2'(1)$	-	Mean number of customers in the orbit during a setup period
$\wp_3'(z)$	-	Mean number of customers in the orbit during a server breakdown
$E(L)$	-	The mean number of customers in the system
$E(N)$	-	The mean length of the queue in the orbit
$E(W)$	-	The expected waiting time of a customer in the orbit
ϖ	-	The steady-state availability of the system
Ω	-	The customer balking rate
C_0	-	Holding cost for each customer in the retrial orbit per unit time
C_μ	-	Cost for maintaining a service rate μ per unit-time
C_θ	-	Cost for providing a retrial rate θ per unit-time
C_α	-	Cost for providing a setup rate per unit-time
C_η	-	Cost for providing a repair rate per unit-time
$\Im$	-	The operating cost function.

Model Formulation

This model considers a single-server queueing system with a retrial orbit. Customer arrivals follow a Poisson process with rate λ . An arriving customer receives immediate service if the server is idle; otherwise, it joins a retrial orbit. The time between successive retrials is exponentially distributed with a rateμ. To conserve energy, the server is deactivated during idle periods. A dormant server is reactivated only by a new primary customer arrival, a process that requires a setup time, exponentially distributed with a rateα. The customer initiating this setup is not served immediately but is instead directed to the retrial orbit. The server is unreliable and may break down during service periods according to a Poisson process with a rateξ. Upon failure, repairs are initiated immediately, with the repair time being exponentially distributed with a rateη. Furthermore, the arrival of a negative customer during a service period causes an immediate server breakdown and the removal of the customer from service. All stochastic processes—inter-arrival times, service times, retrial times, setup times, breakdown times, and repair times—are assumed to be mutually independent.

Steady–State Performance Analysis

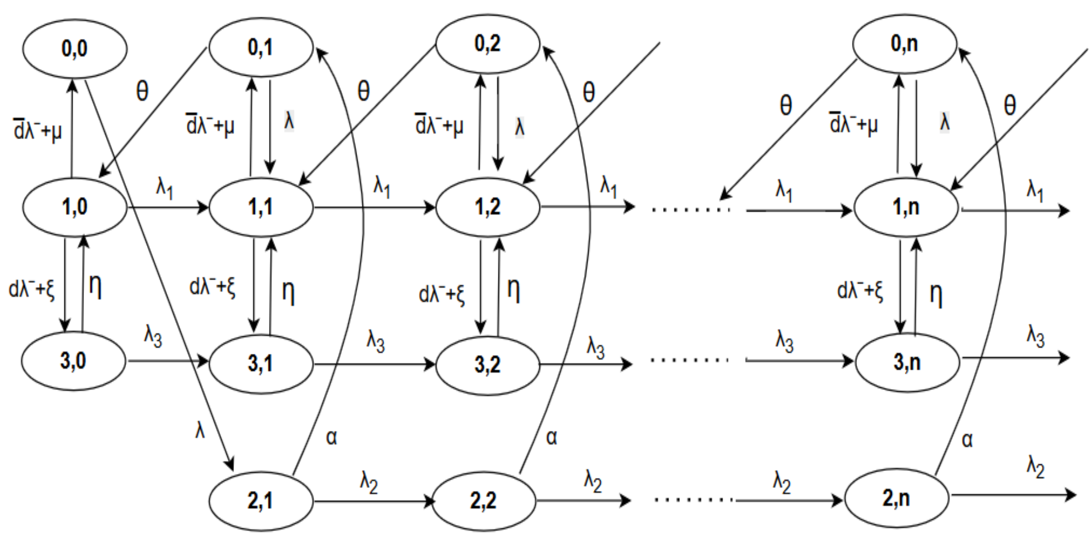
Let $N(t)$ denote the number of customers in the retrial orbit at time$t$, and let $I(t)$ represent the state of the server at time$t$, as defined below.

$$I(t)=\begin{cases}0,\text{ifserverisidle},\\1,\text{if serverisbusy},\\2,\text{ifserverisinsetuptimes},\\3,\text{ifserverisunderrepair}.\end{cases}$$

We know that $X(t)=\{(N(t),I(t));t\geq 0\}$represents a Markov chain. $S=\{(i,j);n\geq 0,i=0,1,2,3\}$ is the state space.

The model incorporates the potential for customer arrivals to join the system during the idle period. The joining probability is state-dependent; specifically, in state $i(i=1,2,3)$, a customer joins with probabilityq_i. Consequently, the effective arrival rate becomes$\lambda_i=\lambda q_i$satisfying$\lambda_i\leq\lambda$. The structure of the underlying continuous-time Markov chain is depicted in the rate transition diagram of figure 1.

Figure 1. **Rate transition diagram**



The partial generating function corresponding to these states is defined as:

$$\wp_0(z)=\sum_{k=0}^{\infty}z^k\Lambda(0,k);$$

$$\wp_1(z)=\sum_{k=0}^{\infty}z^k\Lambda(1,k);$$

$$\wp_2(z)=\sum_{k=1}^{\infty}z^k\Lambda(2,k);$$

$$\wp_3(z) = \sum_{k=0}^{\infty} z^k \Lambda(3,k).$$

The balance equations governing the steady-state probability distribution, under the condition of system stability equation (1) to (9),

$$\Lambda(i,j) = \lim_{t\to\infty} \Lambda\{N(t) = i, I(t) = j\}, (i,j) \in S.$$

$$\lambda\Lambda(0,0) = (\overline{d}\lambda^- + \mu)\Lambda(1,0) \quad (1)$$

$$(\lambda + \theta)\Lambda(0,j) = (\overline{d}\lambda^- + \mu)\Lambda(1,j) + \alpha\Lambda(2,j), j \geq 1 \quad (2)$$

$$(\overline{d}\lambda^- + \mu + d\lambda^- + \xi + \lambda_1)\Lambda(1,j) = \theta\Lambda(0,j+1) + \eta\Lambda(3,j) + \lambda\Lambda(0,j) + \lambda_1\Lambda(1,j-1), j \geq 1 \quad (3)$$

$$(\alpha + \lambda_2)\Lambda(2,1) = \lambda\Lambda(0,0) \quad (4)$$

$$(\alpha + \lambda_2)\Lambda(2,j) = \lambda_2\Lambda(2,j-1), j \geq 2 \quad (5)$$

$$(\eta + \lambda_3)\Lambda(3,0) = (d\lambda^- + \xi)\Lambda(1,0) \quad (6)$$

$$(\eta + \lambda_3)\Lambda(3,j) = (d\lambda^- + \xi)\Lambda(1,j) + \lambda_3\Lambda(3,j-1), j \geq 1. \quad (7)$$

Multiplying the above equations by z^n and summing up which yields,

$$(\lambda+\theta)\wp_0(z) = \alpha\wp_2(z) + (\overline{d}\lambda^- + \mu)\wp_1(z) + \theta\Lambda(0,0) \quad (8)$$

$$(\lambda^- + \mu + \xi + \lambda_1(1-z))\wp_1(z) = \eta\wp_3(z) + \left(\lambda + \frac{\theta}{z}\right)\left[\wp_0(z) - \Lambda(0,0)\right] \quad (9)$$

$$[\alpha + \lambda_2(1-z)]\wp_2(z) = \lambda z\Lambda(0,0) \quad (10)$$

$$\wp_3(z) = \frac{(d\lambda^- + \xi)\wp_1(z)}{\eta + \lambda_3(1-z)}. \quad (11)$$

From equation (10) we get,

$$\Rightarrow \wp_2(z) = \frac{z\lambda\Lambda(0,0)}{\alpha + \lambda_2(1-z)}. \quad (12)$$

From equation (9) we get,

$$\Rightarrow \wp_1(z) = \left\{\frac{\eta(d\lambda^- + \xi)\wp_1(z)}{\eta + \lambda_3(1-z)} + \left(\lambda + \frac{\theta}{z}\right)\left[\wp_0(z) - \Lambda(0,0)\right]\right\}\frac{1}{\lambda^- + \mu + \xi + \lambda_1(1-z)} \quad (13)$$

$$\wp_1(z) = \frac{F_2(z)}{F_1(z)}\left[\wp_0(z) - \Lambda(0,0)\right] \quad (14)$$

$$F_2(z) = \left(\frac{\theta}{z} + \lambda\right) \quad (15)$$

$$F_1(z) = \lambda^- + \xi + \mu + \lambda_1(1-z) - \frac{\eta(d\lambda^- + \xi)}{\eta + \lambda_3(1-z)} \quad (16)$$

$$\wp_0(z)=\frac{F_1(z)}{F_2(z)}\wp_1(z)+\Lambda(0,0). \tag{17}$$

From equation (8), we get the steady-state probabilities of the server being in an idle and busy state are given by,

$$\wp_0(z)=\left\{\frac{\theta F_1(z)\left[\alpha+\lambda_2(1-z)\right]-\left(\bar{d}\lambda^-+\mu\right)\left[\alpha+\lambda_2(1-z)\right]F_2(z)+\alpha z\lambda F_1(z)}{\left[\alpha+\lambda_2(1-z)\right]\left[(\lambda+\theta)F_1(z)-\left(\bar{d}\lambda^-+\mu\right)F_2(z)\right]}\right\}\Lambda(0,0) \tag{18}$$

$$\wp_1(z)=\frac{F_2(z)}{F_1(z)}\left\{\frac{\theta F_1(z)\left[\alpha+\lambda_2(1-z)\right]-\left(\bar{d}\lambda^-+\mu\right)\left[\alpha+\lambda_2(1-z)\right]F_2(z)+\alpha z\lambda F_1(z)}{\left[\alpha+\lambda_2(1-z)\right]\left[(\lambda+\theta)F_1(z)-\left(\bar{d}\lambda^-+\mu\right)F_2(z)\right]}-1\right\}\Lambda(0,0). \tag{19}$$

Substituting $z=1$in (15) and (16), we get,

$$F_2(1)=\theta+\lambda, F_1(1)=\bar{d}\lambda^-+\mu.$$

Substituting $z=1$in equation (18) we obtain,

$$F_1{}'(z)=-\lambda_1-\frac{\lambda_3\eta(d\lambda^-+\xi)}{[\eta+\lambda_3(1-z)]^2}, F_2{}'(z)=\frac{-\theta}{z^2}$$

$$\wp_0(1)=\Lambda(0,0)\left\{\frac{\theta F_1(1)\alpha+\alpha\lambda F_1(1)-\left(\bar{d}\lambda^-+\mu\right)\alpha F_2(1)}{\alpha\left[(\lambda+\theta)F_1(1)-\left(\bar{d}\lambda^-+\mu\right)F_2(1)\right]}\right\}=\frac{0}{0}.$$

$F_1{}'(1)=-\lambda_1-\frac{\lambda_3(d\lambda^-+\xi)}{\eta}, F_2{}'(1)=-\theta.$

By L' Hospital's rule we get,

$$\wp_0(1)=\frac{(\theta+\lambda)\left\{-\alpha\left[\eta\lambda_1+\lambda_3\left(d\lambda^-+\xi\right)\right]+\eta\lambda_2\left(\bar{d}\lambda^-+\mu\right)+\alpha\eta\left(\bar{d}\lambda^-+\mu\right)\right\}-\theta\lambda_2\eta\left(\bar{d}\lambda^-+\mu\right)}{\alpha\left\{-(\theta+\lambda)\left[\lambda_1\eta+\lambda_3\left(d\lambda^-+\xi\right)\right]+\eta\theta\left(\bar{d}\lambda^-+\mu\right)\right\}}\Lambda(0,0) \tag{20}$$

$$\wp_1(z)=\frac{F_2(z)}{F_1(z)}\left[\wp_0(z)-\Lambda(0,0)\right]$$

$$\wp_1(1)=\frac{(\theta+\lambda)\lambda\eta(\alpha+\lambda_2)}{\alpha\left\{-(\theta+\lambda)\left[\lambda_1\eta+\lambda_3\left(d\lambda^-+\xi\right)\right]+\eta\theta\left(\bar{d}\lambda^-+\mu\right)\right\}}\Lambda(0,0). \tag{21}$$

The steady-state probability associated with the server's setup state is obtained from equation (12) by substituting$z=1$,

$$\Rightarrow\wp_2(1)=\frac{\lambda}{\alpha}\Lambda(0,0). \tag{22}$$

Substituting the result from equation (21) and evaluating at $z=1$in Equation (11) yields the steady-state probability for the server's repair state,

$$\wp_3(1)=\frac{\lambda(\alpha+\lambda_2)(d\lambda^-+\xi)(\theta+\lambda)}{\alpha\left\{-(\theta+\lambda)\left[\lambda_1\eta+\lambda_3(d\lambda^-+\xi)\right]+\eta\theta(\bar{d}\lambda^-+\mu)\right\}}\Lambda(0,0). \quad (23)$$

We know that, $\wp_0(1)+\wp_1(1)+\wp_2(1)+\wp_3(1)=1.$

Adding the equations (20), (21), (22), (23) and substituting z=1 we get,

$$\Lambda(0,0)=\frac{\alpha\{\eta\theta(\bar{d}\lambda^-+\mu)-(\theta+\lambda)[\eta\lambda_1+\lambda_3(\xi+d\lambda^-)]\}}{(\theta+\lambda)\{\lambda[\eta+d\lambda^-+\xi](\lambda_2+\alpha)+\alpha\eta(\bar{d}\lambda^-+\mu)-(\alpha+\lambda)[\lambda_3(\xi+d\lambda^-)+\eta\lambda_1]\}+(\theta+\lambda_2)\lambda\eta(\bar{d}\lambda^-+\mu)}. \quad (24)$$

Let, $\wp_2'(1)$ denote the mean number of customers in the orbit during a setup period. This measure is obtained as follows:

Differentiating equation (12) w.r.t z and substituting z=1 we get,

$$\wp_2'(1)=\frac{\lambda(\alpha+\lambda_2)}{\alpha^2}\Lambda(0,0). \quad (25)$$

Let $\wp_3'(z)$represent the mean number of customers in the orbit during a server breakdown. This measure is obtained as follows:

Differentiating (11) w.r.t z and substituting z =1 we obtain,

$$\wp_3'(1)=(d\lambda^-+\xi)\left\{\frac{\wp_1'(1)}{\eta}+\frac{\wp_1(1)\lambda_3}{\eta^2}\right\}. \quad (26)$$

Let $\wp_0'(1)$ denote the mean number of customers in the orbit during an idle period. This measure is obtained as follows:

The mean queue length of the orbit in an idle period is $\wp_0'(1)$such that,

$$\wp_0'(1)=\frac{(\bar{d}\lambda^-+\mu)}{(\theta+\lambda)}\wp_1'(1)+\frac{\lambda(\alpha+\lambda_2)}{\alpha(\theta+\lambda)}\Lambda(0,0). \quad (27)$$

Differentiating equation (18) and applying L'Hospital's rule, we get,

$$\wp_0'(1)=\Lambda(0,0)\frac{\alpha[\tau u+M\theta]\left\{[2u(\alpha\lambda-\theta\lambda_2)-2(\lambda_2+\alpha)M\theta]+2\lambda_2\left[\tau u+\tau M+\frac{M\lambda\lambda_2}{\alpha}\right]\right\}-\alpha\tau vM\lambda(\alpha+\lambda_2)+2\alpha M\theta[\alpha\tau u+\tau M\alpha+\lambda_2 M\lambda]}{2[\tau u+M\theta]^2\alpha^2}. \quad (28)$$

Were

$$F_1(1)=\bar{d}\lambda^-+\mu=M, F_2(1)=(\lambda+\theta)=\tau$$

$$F_1'(1)=-\lambda_1-\frac{\lambda_3(d\lambda^-+\xi)}{\eta}=u(say), F_2'(1)=-\theta$$

$$F_1''(1)=\frac{-2{\lambda_3}^2(d\lambda^-+\xi)}{\eta^2}=v(say), F_2''(1)=2\theta$$

$$\wp_0'(1)=\Lambda(0,0)\frac{X\{2[\alpha u(\alpha\lambda-\lambda_2\theta)-M\alpha\theta(\lambda_2+\alpha)]+2\lambda_2 Y\}-\alpha\tau vM\lambda(\alpha+\lambda_2)+2\alpha M\theta Y}{2\alpha^2X^2}. \quad (29)$$

Were

$$X=\tau u+M\theta, Y=\alpha\tau u+M\alpha\tau+M\lambda_2\lambda.$$

Let $\wp_1'(1)$denote the mean number of customers in the orbit during a busy period. This measure is obtained as follows:

$$\wp_1'(1)=\Lambda(0,0)\left\{-\frac{\lambda(\lambda_2+\alpha)}{M\alpha}+\frac{\tau\left\{[u(\alpha\lambda-\theta\lambda_2)-(\lambda_2+\alpha)M\theta]+\left[\tau u+M\tau+\frac{M\lambda_2\lambda}{\alpha}\right]\lambda_2\right\}}{\alpha[\tau u+\theta M]M}-\frac{\tau^2 v\lambda(\lambda_2+\alpha)+2\theta\tau[\alpha\tau u+M\alpha\tau+M\lambda_2\lambda]}{2(\alpha[\tau u+M\theta]^2)}\right\}. \quad (30)$$

$$\text{Let}K=\frac{X\{2[\alpha u(\alpha\lambda-\lambda_2\theta)-M\alpha\theta(\lambda_2+\alpha)]+2\lambda_2Y\}-\alpha\tau vM\lambda(\alpha+\lambda_2)+2\alpha M\theta Y}{2\alpha^2X^2}$$

$$\left.\begin{aligned}
&\wp_0'(1)=\Lambda(0,0).K \text{ and } \wp_1(1)=\frac{\tau\lambda\eta(\alpha+\lambda_2)}{\alpha\left\{-\tau\left[\lambda_1\eta+\lambda_3\left(d\lambda^-+\xi\right)\right]+\eta\theta M\right\}}\Lambda(0,0),\\
&\wp_1'(1)=\left\{-\frac{(\alpha+\lambda_2)\lambda}{M\alpha}+\frac{\tau}{M}K\right\}\Lambda(0,0),\\
&\wp_2'(1)=\Lambda(0,0).\frac{\lambda(\alpha+\lambda_2)}{\alpha^2},\\
&\wp_3'(1)=\left(d\lambda^-+\xi\right)\left\{\frac{\wp_1'(1)}{\eta}+\frac{\wp_1(1)\lambda_3}{\eta^2}\right\}.
\end{aligned}\right\} \tag{31}$$

Therefore, the mean length of the queue in the orbit is obtained as,

$$E(N)=\wp_0'(1)+\wp_1'(1)+\wp_2'(1)+\wp_3'(1).$$

Substituting equation (31) we get,

$$E(N)=\left[K\left(\frac{\tau}{M}+1\right)-\frac{(\lambda_2+\alpha)\lambda}{\alpha}\left[\frac{1}{M}-\frac{1}{\alpha}\right]+(d\lambda^-+\xi)\left\{\frac{1}{\eta}C+\frac{\lambda_3}{\eta^2}D\right\}\right]\Lambda(0,0). \tag{32}$$

Were

$$C=-\frac{\lambda(\alpha+\lambda_2)}{\alpha M}+\frac{\tau}{M}K and D=\frac{\tau\lambda\eta(\alpha+\lambda_2)}{\alpha\{-\tau[\lambda_1\eta+\lambda_3(d\lambda^-+\xi)]+\eta\theta M\}}.$$

The mean number of customers in the system $E(L)$ is the sum of the mean orbit length and the server's utilization factor (the probability that the server is busy).

$$E(L)=E(N)+D.\left[1+\frac{(d\lambda^-+\xi)}{\eta}\right]\Lambda(0,0).$$

$$\begin{aligned}
&E(L)=\wp_0'(1)+\wp_1'(1)+\wp_2'(1)+\wp_3'(1)+\wp_1(1)+\wp_3(1)\\
&E(L)=E(N)+\wp_1+\wp_3
\end{aligned} \tag{33}$$

Let $E(W)$ denote the expected waiting time of a customer in the orbit. This performance measure is obtained as follows:

$$E(W)=\frac{E(N)}{\lambda_k}, where \lambda_k=\lambda_1P_1+\lambda_2P_2+\lambda_3P_3. \tag{34}$$

$$E(W)=\frac{\Lambda(0,0)}{\lambda_k}\left[K\left(1+\frac{\tau}{M}\right)+\frac{\lambda(\alpha+\lambda_2)}{\alpha}\left[\frac{1}{\alpha}-\frac{1}{M}\right]+(d\lambda^-+\xi)\left\{\frac{1}{\eta}C+\frac{\lambda_3}{\eta^2}D\right\}\right]. \tag{35}$$

The total arrival rate in the retrial orbit is λ_k.

The steady-state availability of the system is computed as,

$$\varpi=\wp_0(1)+\wp_1(1)+\wp_2(1).$$

$$\varpi=\Lambda(0,0)\left\{\begin{aligned}&\frac{(\theta+\lambda)\{-\alpha[\eta\lambda_1+\lambda_3(d\lambda^-+\xi)]+\eta\lambda_2(\bar{d}\lambda^-+\mu)+\alpha\eta(\bar{d}\lambda^-+\mu)\}-\theta\lambda_2\eta(\bar{d}\lambda^-+\mu)}{\alpha\{-(\theta+\lambda)[\lambda_1\eta+\lambda_3(d\lambda^-+\xi)]+\eta\theta(\bar{d}\lambda^-+\mu)\}}\\&+\frac{(\theta+\lambda)\lambda\eta(\alpha+\lambda_2)}{\alpha\{-(\theta+\lambda)[\lambda_1\eta+\lambda_3(d\lambda^-+\xi)]+\eta\theta(\bar{d}\lambda^-+\mu)\}}+\frac{\lambda}{\alpha}\end{aligned}\right\}. \tag{36}$$

The customer balking rate is calculated as,

$$\Omega=(\lambda-\lambda_1)\wp_1(1)+(\lambda-\lambda_2)\wp_2(1)+(\lambda-\lambda_3)\wp_3(1). \tag{37}$$

NUMERICAL ANALYSIS

The sensitivity of system performance and cost is analysed graphically, quantifying the impact of key parameters. All numerical experiments are conducted with parameter sets that fulfil the system's stability requirements. The interpretations presented herein are intended to provide a conceptual linkage between the proposed queueing framework and neural information processing and cognitive systems, rather than to establish a direct biological model.

Numerical Analysis of Performance Measures

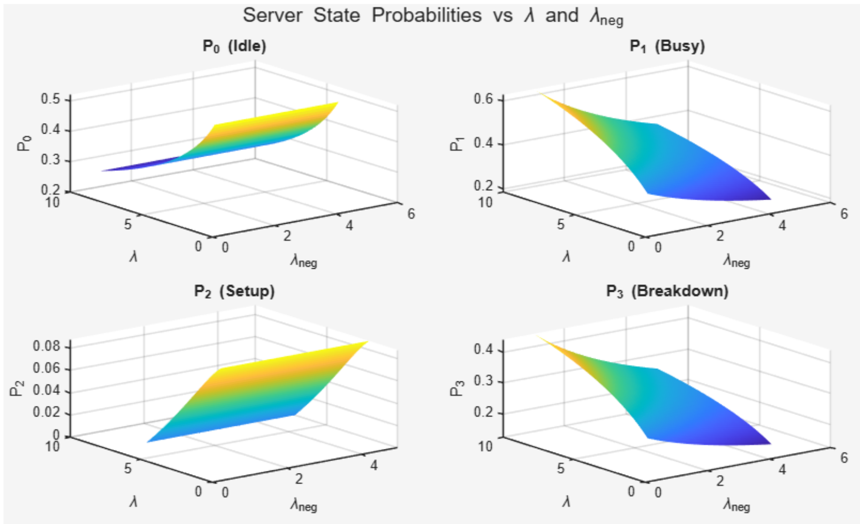
Examples 1:

The Server state probability vs. λ. Calculating $\wp_0, \wp_1, \wp_2, \wp_3$ using equations (20), (21), (22), and (23), for the cases given below,

Case (a):

We get the below graph, when $\eta = 5, \mu = 6, \alpha = 4, \theta = 3, \xi = 2, \lambda_1 = 1.5, \lambda_2 = 2, \lambda_3 = 1, d\lambda^- = 1.5, \overline{d}\lambda^- = 1.5.$

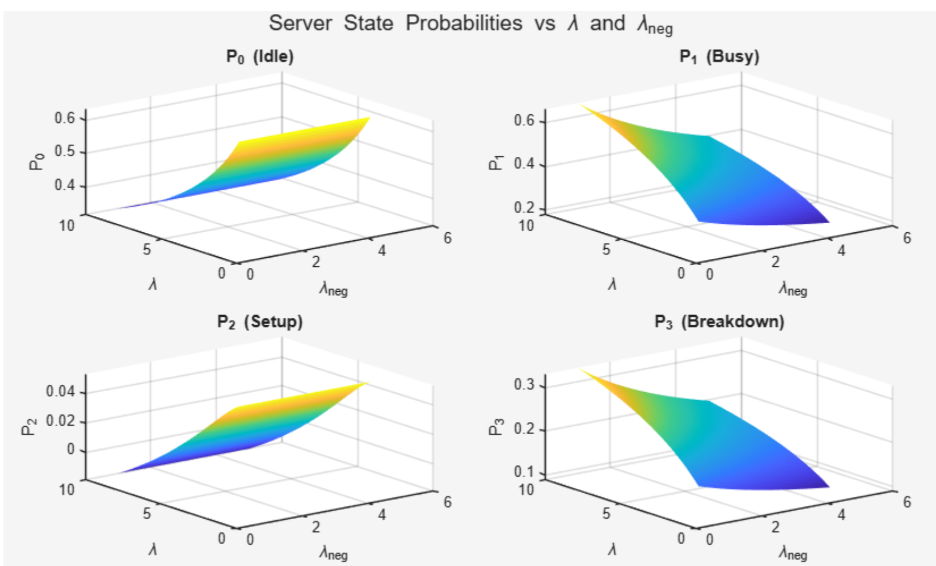
Figure 2. **Case 1(a) gives 3D plot for server state probability**



Case (b):

We get the below graph, when $\eta = 8, \mu = 7, \alpha = 7, \theta = 2, \xi = 2, \lambda_1 = 1.3, \lambda_2 = 2.5, \lambda_3 = 1, d\lambda^- = 2, \overline{d}\lambda^- = 1.5.$

Figure 3. **Case 1(b) gives 3D plot for server state probability**



The numerical analysis demonstrates the capability of the proposed model to characterize the behavioural patterns of the server states across a range of parameter settings (Figure 2 and Figure 3). The following analysis identifies key deviations:

- At low arrival rates (λ), the model indicates a relatively slower transition from idle to active operation, resulting in higher idle-state probabilities and lower busy-state probabilities.
- As λ increases, the setup-state probability demonstrates increased responsiveness to parameter variation.
- In contrast, the breakdown-state probability remains comparatively less sensitive to variations inλacross the considered parameter range.
- From the perspective of neural information processing and cognitive systems, lower arrival intensity may represent reduced sensory stimulation or cognitive demand, resulting in prolonged inactive processing states and delayed cognitive engagement. As information inflow increases, cognitive preparation and activation become more responsive; however, the model indicates that interruptions associated with cognitive fatigue or transient processing disruptions may still not be fully captured under higher processing loads.

Example 2:

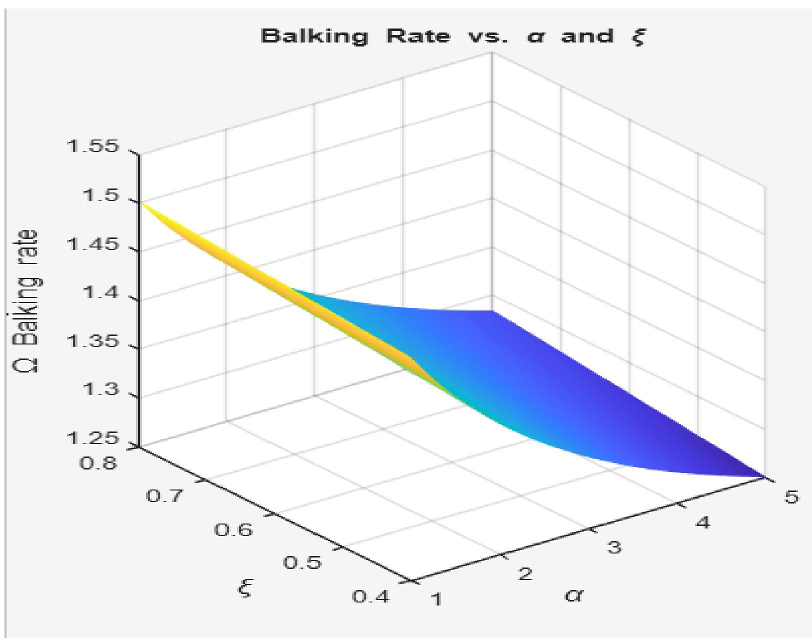
We plot the behavior of balking rate againstλ, provided by equations (21), (22), (23), (24), and (37) were,

$$\Omega = (\lambda - \lambda_1)\wp_1(1) + (\lambda - \lambda_2)\wp_2(1) + (\lambda - \lambda_3)\wp_3(1).$$

Case (a):

When $\lambda = 4, \lambda_1 = 1.5, \lambda_2 = 2, \lambda_3 = 1, \mu = 7, \theta = 5, \eta = 6, d\lambda^- = 1.5, \overline{d}\lambda^- = 2.$

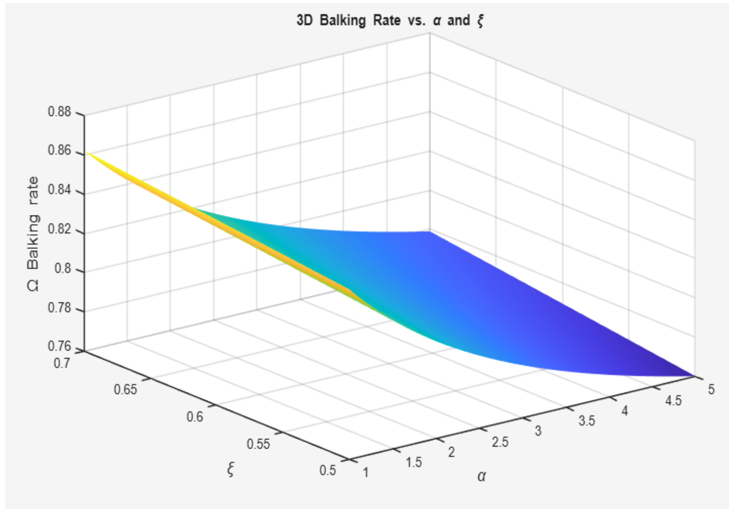
Figure 4. **Case 2(a) behavior of balking rate against**λ**when** $\xi = 0.4, 0.6, 0.8$.



Case (b):

When $\lambda = 3, \lambda_1 = 1.5, \lambda_2 = 2, \lambda_3 = 1, \mu = 7, \theta = 3, \eta = 4, d\lambda^- = 1.5, \overline{d}\lambda^- = 2.$

Figure 5. **Case 2(b) behavior of balking rate against λ when $\xi = 0.5, 0.6, 0.7$.**



The analysis of system's operation is defined by service and transition rates that is a factor of balking behaviour (Figure 4 and Figure 5). Following patterns of distinct response are noticed:

- High-Intensity Regime: Increased α leads to rapid deterioration in the balking rate, with system performance of high sensitivity to varying ξ levels.
- Low-Intensity Regime: The decrease in balking with increasing α is more precise. The marginal effect of ξ is moderated, consistent with a regime of reduced gain and steady-state behaviour.
- From the perspective of neural information processing and cognitive systems, this behavior may conceptually represent adaptive cognitive resource allocation in response to varying information. Under conditions of high information inflow, faster cognitive activation and attention allocation enhance processing readiness and reduce information rejection. In contrast, under lower information demand, processing resources remain balanced and stable, resulting in reduced sensitivity to activation dynamics and sustained cognitive performance.

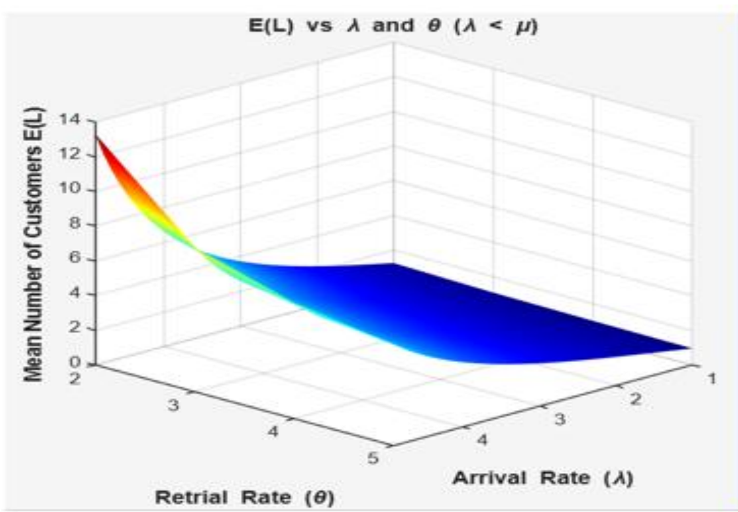
Example 3

The mean number of customers in the system E(L) represented by equation (33) is derived from equations (24) and (32).

Case (a):

When $\lambda_1 = 0.8, \lambda_2 = 1.2, \lambda_3 = 0.9, \alpha = 3, \eta = 3.5, \xi = 0.6, \mu = 5, d = 0.6, \lambda^- = 0.5$.

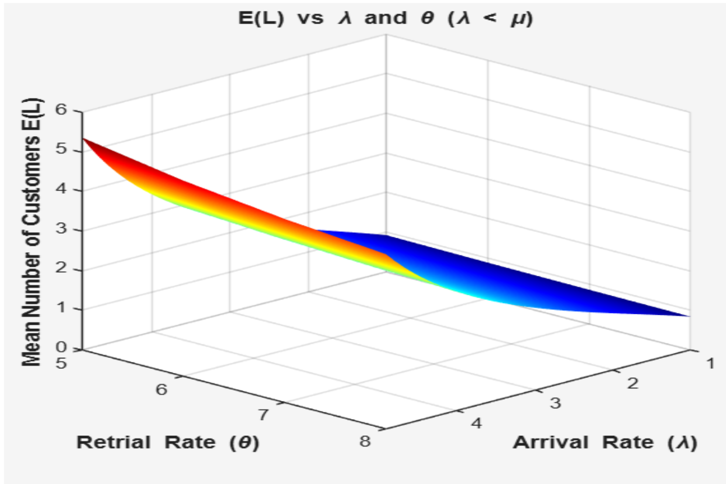
Figure 6. **Case 3(a) mean number of customers in the service vs. the arrival rate for different retail rates,** $\theta = 2,3,4,5$.



Case (b):

When $\lambda_1 = 0.5, \lambda_2 = 1.1, \lambda_3 = 0.8, \alpha = 3, \eta = 3, \xi = 0.6, \mu = 5, d = 0.6, \lambda^- = 0.5$.

Figure 7. **Case 3(b) mean number of customers in the service vs. the arrival rate for different retail rates,** $\theta = 5,6,7,8$.



The graphs (Figure 6 and Figure 7) show the association of the mean number of customers in the system, E(L), on the arrival rate (λ) and retrial rate (θ). The metric E(L) demonstrates a positive relationship with λ, rising as the increased arrival intensity elevates system congestion. In contrast, E(L) decreases with θ, as an accelerated retrial process mitigates orbit queue build up by re-introducing customers to the server more efficiently.

From the perspective of neural information processing and cognitive systems, the arrival rate may represent the intensity of incoming information entering the cognitive processor. Higher arrival intensity leads to increased information accumulation and cognitive workload, reducing processing efficiency. In contrast, a higher retrial rate may represent repeated information retrieval, improve information access and reduce unresolved cognitive load.

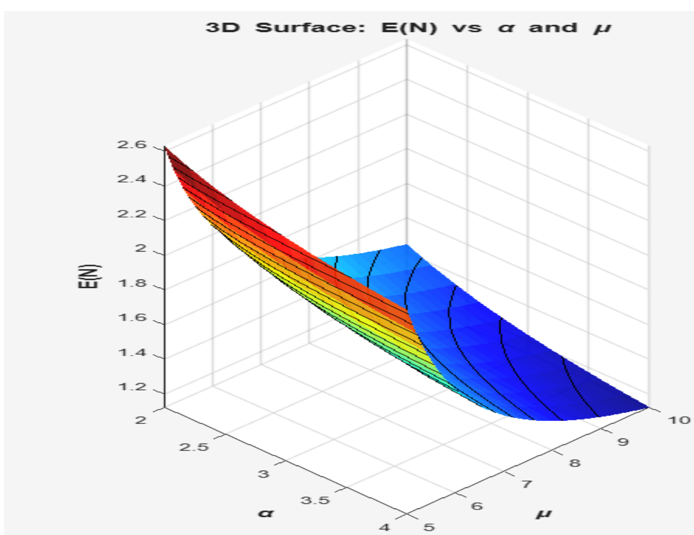
Example 4.

The behaviour of the mean orbit queue length, $E(N)$ is illustrated against varying values of μ and α. This result is derived from the analytical expressions in equations (24) and (32)

Case (a):

When $\lambda = 2, \lambda_1 = 1.5, \lambda_2 = 1.2, \lambda_3 = 0.8, \theta = 2.5, \eta = 4, \xi = 1, \lambda^- = 0.4$.

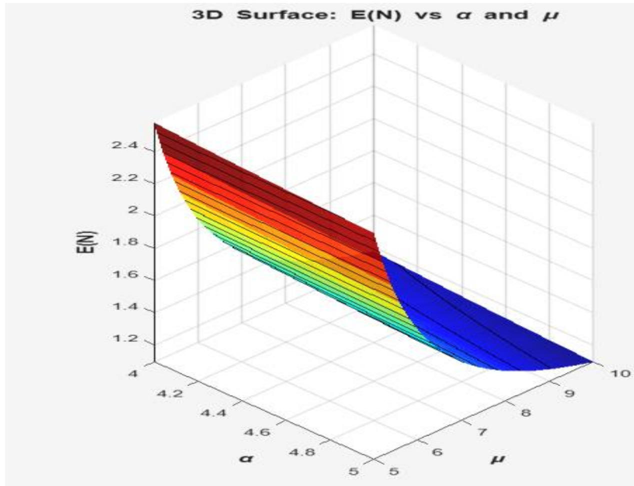
Figure 8. **Case 4 (a) mean queue length of the orbit vs.** μ**for different values of** $\alpha\alpha = 2.0,2.5,3.0,3.5$.



Case (b):

When $\lambda = 2.1, \lambda_1 = 1.7, \lambda_2 = 1.5, \lambda_3 = 0.9, \theta = 3.0, \eta = 4.5, \xi = 1.1, \lambda^- = 0.4$.

Figure 9. **Case 4 (b) mean queue length of the orbit vs. $\mu$for different αvalues $\alpha = 4.0,4.2,4.4,4.6$.**



The mean orbit queue length E(N) decreases with the service rate (μ) but increases as the setup rate

(α) decreases. This occurs because a high μ enables rapid customer service, draining the orbit, while a low αcauses prolonged setup times, resulting in customer pile-up and a longer orbit queue.

From the perspective of neural information processing and cognitive systems, a higher service rate (μ) may represent faster cognitive processing and information handling, which reduces the accumulation of unresolved information and lowers cognitive workload. In contrast, a lower setup rate (α) may represent slower cognitive activation or delayed attention allocation, leading to increased information buildup and reduced processing efficiency.

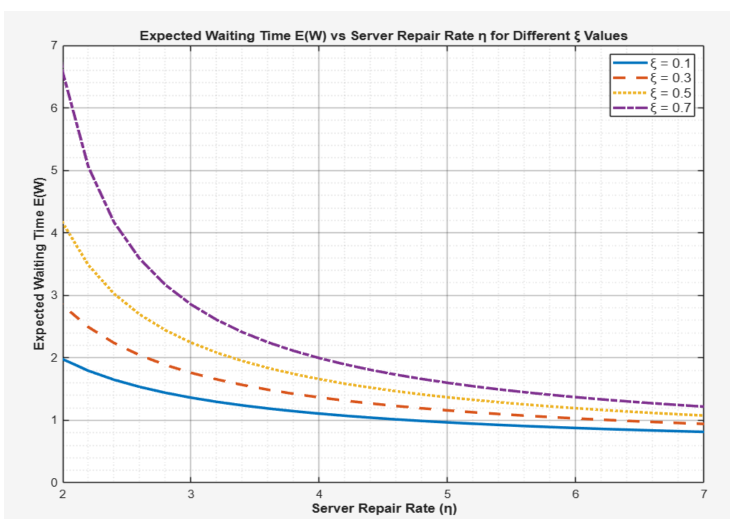
Example 5:

The expected waiting time EW) vs.η. The equations (24) and (35) are used to find the relation between E (W), η for varyingξ.

Case (a):

When $\lambda = 2.5, \lambda_1 = 1.5, \lambda_2 = 0.6, \lambda_3 = 0.2, \alpha = 1.2, \mu = 4.0, \theta = 1.0, \lambda^- = 1.0$.

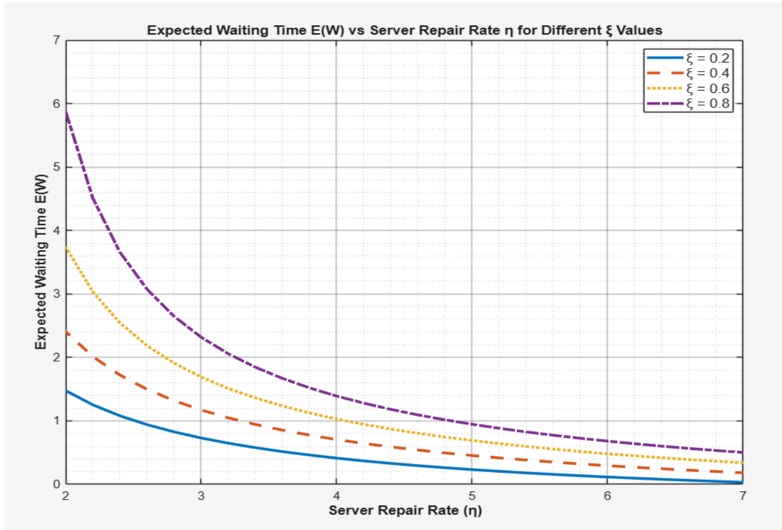
Figure 10. **Case 5(a) expected waiting time E (W) vs. server repair rate η for differentξ, $\xi = 0.1,0.3,0.5,0.7$.**



Case (b):

When $\lambda = 3, \lambda_1 = 1.8, \lambda_2 = 1.2, \lambda_3 = 1.0, \alpha = 1.2, \mu = 5.0, \theta = 1.0, \lambda^- = 1.0$.

Figure 11. **Case 5(b) expected waiting time E (W) vs. server repair rate η for different ξ, $\xi = 0.2,0.4,0.6,0.8$.**



The expected waiting time exhibits a negative correlation with the repair rate (η) and a positive correlation with the failure rate (ξ), as shown in the graph. An increased η improves the availability by minimizing repair downtime, hence reducing the waiting times. Conversely, an increased ξ suppresses the availability of the server, which leads to a sustained waiting period of the customers.

From the perspective of neural information processing and cognitive systems, the expected waiting time may represent the delay in information processing or retrieval. A higher repair rate (η) may indicate faster recovery of processing resources, thereby reducing processing delays and improving system efficiency. In contrast, a higher failure rate (ξ) may represent cognitive interruptions or temporary processing disruptions, leading to longer waiting times and reduced processing performance.

Cost Analysis

The minimum cost is obtained by optimizing the cost. The cost parameters are as follows,

C_0 = Holding cost for each customer in the retrial orbit per unit time.

C_μ = Cost for maintaining a service rate μper unit-time.

C_θ = Cost for providing a retrial rateθper unit-time.

C_α = Cost for providing a setup rateαper unit-time.

C_η = Cost for providing a repair rateηper unit-time.

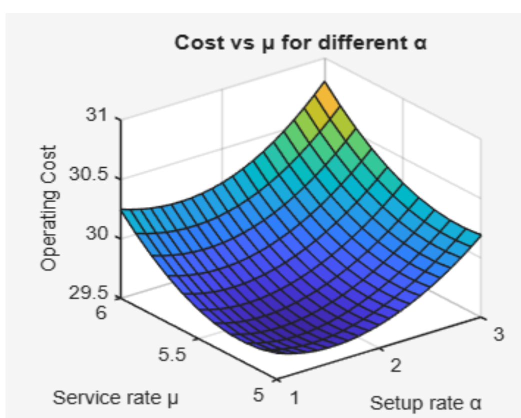
Therefore, the operating cost function is given by

$$\Im = C_0E(N) + C_\mu\mu + C_\theta\theta + C_\alpha\alpha + C_\eta\eta \qquad (38)$$

Example 6:

An optimal service rate at which the total operating cost is minimized, by balancing the choice between waiting cost and setup cost.

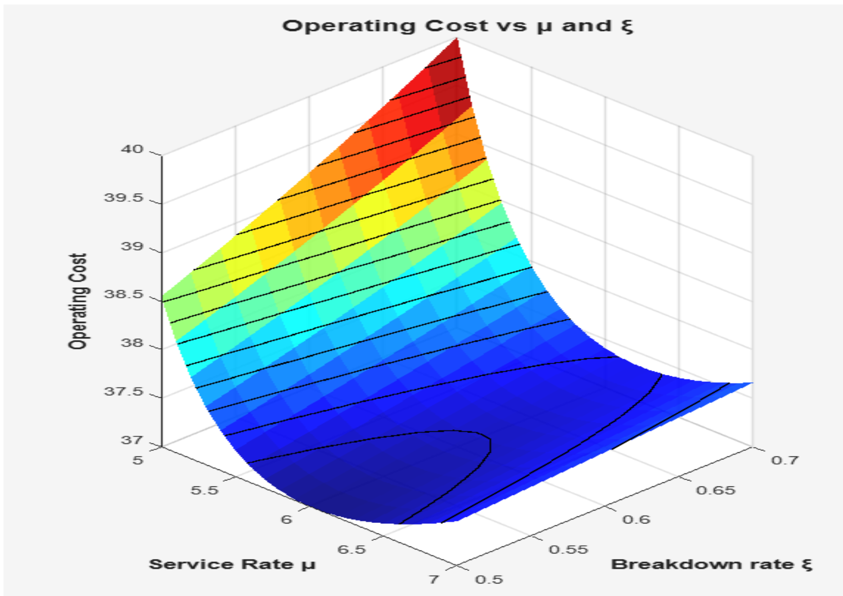
Figure 12. **Operation cost vs. service rate$\mu$for different setup rateα**



When the service rate (μ) increases, the total operating cost first decreases and then increases. This is shown in the graph (Figure 12) as a relationship between the service rate and the system cost and represented as U-shaped. Customer waiting time is reduced at lower value of μ, henceforth increasing the service rate allows the server to set up more quickly and hence lowering the waiting cost. At the same time, the reduction in waiting cost increases the setup cost, which leads to a decrease in total cost.

Nevertheless, as μ continues to increase above a critical value, the setup cost is predominant**,** and the total operating cost starts to increase. This is because a higher setup rate requires more energy, making the system more expensive despite the reduced waiting time.

Figure 13. **Operation cost vs. service rate$\mu$for different breakdown rateξ**



The system demonstrates an irregular dependence of operating costs onξ. The trend reveals that the cost reduces initially; however, later a sharp rise in costs occurs, which represents increase in server breakdowns, constituting the major cost factor.

This graph (Figure 13) demonstrates that there exists an optimal service rate where the total operating cost is minimized, by optimizing the waiting cost and setup cost**.**

From the perspective of neural information processing and cognitive systems, the service rate μ may represent the speed of information processing. Initially, increasing μ reduces processing delay and lowers the overall cost by improving information flow. However, beyond a certain level, additional processing effort increases resource consumption and raises the total cost.

Similarly, the failure rate ξ may represent cognitive interruptions or temporary processing disruptions. Higher failure intensity increases recovery requirements and operational cost. These results indicate the existence of an optimal processing level that balances processing efficiency, resource utilization, and system cost.

SENSITIVITY ANALYSIS OF THE MODEL

This research provides an empirical analysis of system performance, influencing the changes in parameters with a repairable M/M/1 retrial queue, along with setup times and negative customer arrivals. The analysis indicates that an increase in the arrival rate enhances the system, leading to higher congestion and reduced idle and setup time. A higher retrial rate successfully reduces the orbit size and average waiting time, improving efficient service. Optimization in service and setup rates contributes to faster customer handling and reduced delay in the system, while an increased breakdown rate results in more regular service failures and longer waiting periods. On the other hand, a higher repair rate improves the service availability of the system by reducing downtime. Moreover, the cost function, designed by combining the costs associated with holding, service, retrials, setup, and repair, provides a basis for detecting optimal setups that minimize total functional expenses. These results provide valuable factors for system administrators that aim for balanced performance, cost, and reliability—in particular, critical infrastructure such as telecommunications, cloud services, and energy-aware systems—making this study relevant to both academic research and practical utilization in environments with high-performance service.

A systematic examination of parameter sensitivities in a repairable M/M/1 retrial queue is done, involving setup times and negative customer arrivals. The analysis determines that the arrival rate (λ) directly influences system load, where elevated values increase queue lengths and reduce idle intervals. The retrial rate (θ) illustrates an inverse relationship with orbital queue size and mean waiting time, reflecting enhanced service accessibility. Similarly, accelerated service (μ) and setup (α) rates collectively minimise the system delay through fast execution. According to system dependability, the repair rate (η) makes up for the performance by restoring operational capacity, when the failure rate (ξ) is directly correlated with increased service disruption and waiting time. Operating conditions are made possible by a cost-benefit framework that includes service, retrial, setup, and maintenance costs. The key findings provided by both theoretical contributions to queueing theory and practical methodologies for engineering design in reliability-sensitive service environments.

APPLICATIONS TO WIRELESS COMMUNICATION AND NEURAL INFORMATION PROCESSING AND COGNITIVE SYSTEMS

In wireless communication systems, negative customers could be used to model interference factors that reduce the efficiency of compromising transmissions. These interruptions may include signal interference, collisions of data packets, jamming, or unexpected channel failures. When a negative customer arrives in the system, it in addition active connections or removes queued data packets. This results in increased re-transmissions, higher delay, diminished throughput, and loss of information. The presence of negative customers would also affect the stability of the system and reliability, especially in high-traffic. Introducing the behaviour of negative customers into the model, the network enables the designers to evaluate external disturbances impact on communication quality. As a result, the system, in the presence of repetitive disruptions, would be able to maintain performance, which is essential for good service in modern wireless networks.

The proposed repairable M/M/1 retrial queue with setup times and negative arrivals provides a conceptual stochastic framework for analyzing selected aspects of neural information processing and cognitive systems under conditions of interruptions, resource limitations, and energy constraints.

In this interpretation, arriving customers represent incoming sensory stimuli or information requests entering the cognitive processing unit, and the server denotes the available neural or cognitive processing resource. The retrial mechanism captures repeated memory access and information retrieval attempts that occur when information cannot be processed immediately. The setup period represents cognitive preparation, attention allocation, or activation delays before processing begins. Negative arrivals represent interference, forgetting, or inhibitory phenomena that prevent information from being processed successfully. Server breakdowns represent temporary cognitive fatigue, neural disruption, or interruptions in processing activity, whereas the repair processes correspond to recovery and restoration of cognitive resources. The idle-time shutdown policy further reflects energy-efficient behavior by allowing processing resources to remain inactive during periods of low demand and reactivate when new information arrives.

The numerical and cost-optimization analyses provide insight into balancing information-processing efficiency, recovery mechanisms, retrial behavior, and resource utilization. The proposed framework may serve as a useful mathematical abstraction for studying interruption-sensitive, energy-aware, and resource-constrained information processing environments relevant to neural and cognitive systems.

CONCLUSION

This study develops an analytical framework using repairable M/M/1 retrial queueing system integrating setup times and the disruptive effect of negative customer arrivals. The model interrupts an energy-saving process that allows the server to remain dormant during idle periods and is operational only on the arrival of a new customer, who in turn enters the retrial orbit rather than being served immediately. The system's performance is analysed using the probability generating function method, enabling the derivation of steady-state probabilities and principal performance metrics such as the mean orbit length. A cost model was designed to determine optimal operational parameters, followed by numerical analysis to evaluate the system performance sensitivity under diverse conditions.

Both the behaviour of retrial and failures of server make this model highly significant to real-world applications such as wireless networks and distributed computational architectures, where energy conservation and interrupted service are critical factors.

REFERENCES

Atencia I, Galán-García M Á, Aguilera-Venegas G, Galán-García J L. A non-Markovian retrial queueing system [J]. *Journal of Computational and Applied Mathematics*, 2023, 431: 115277. https://doi.org/10.1016/j.cam.2023.115277

Bagyam J E A, Suganthi P, Visali P. Unreliable single server retrial queueing model with repeated vacation [J]. *International Journal of Mathematics in Operational Research*, 2023, 26(3): 357-372. https://doi.org/10.1504/IJMOR.2023.134838

Bouchentouf A A, Yahiaoui L, Ziad I. Modeling and optimizing an AMS with DV policy, waiting servers, impatient customers, and failures: A queueing analysis [J]. *Results in Control and Optimization*, 2024, 15: 100427. https://doi.org/10.1016/j.rico.2024.100427

Bura G S, Sharma H. Maximum likelihood and Bayesian estimation on M/M/1 queueing model with balking [J]. *Communications in Statistics-Theory and Methods*, 2024, 53(14): 5117-5145. https://doi.org/10.1080/03610926.2023.2208695

Cao S, Liu Y. Integrating queueing network and ACT-R cognitive architectures [C] // *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 2011, 55(1): 836-840. https://doi.org/10.1177/1071181311551174

Chen F M, Liu T H, Ke J C. On an unreliable-server retrial queue with customer feedback and impatience [J]. *Computers & Industrial Engineering*, 2024, 191: 110145.

Das D, Tamuli A, Choudhury A. Classical and Bayesian estimation of performance measures in a novel M/M/1 queueing model with balking [J]. *Communications in Statistics-Simulation and Computation*, 2026, 55(6): 2281-2298. https://doi.org/10.1080/03610918.2025.2464075

Dhibar S, Jain M. Metaheuristics and strategic behavior of Markovian retrial queue under breakdown, vacation and Bernoulli feedback [J]. *Applied Intelligence*, 2025, 55(4): 273. https://doi.org/10.1007/s10489-024-05978-x

Hablal H, Touche N, Alem L, Bouchentouf A A, Boualem M. Lower and upper stochastic bounds for the joint stationary distribution of a non-preemptive priority retrial queueing system [J]. *Hacettepe Journal of Mathematics and Statistics*, 2023, 52(5): 1438-1460. https://doi.org/10.15672/hujms.1183966

Jayamani V N, Madheswari S P, Suganthi P, Josephine S A. An N-policy M/M/1 queueing model for energy saving mechanism in networks [J]. *Sustainable Computing: Informatics and Systems*, 2024, 44: 101016. https://doi.org/10.1016/j.suscom.2024.101016

Jeganathan K, Selvakumar S, Srinivasan K, Anbazhagan N, Joshi G P, Cho W. Two-types of service facilities in interconnected stochastic queueing and queueing-inventory system with marked Markovian arrival process [J]. *Ain Shams Engineering Journal*, 2024, 15(9): 102963. https://doi.org/10.1016/j.asej.2024.102963

Kalaiselvi J, Saravanarajan M C. Exploring non-Markovian dynamics: Stochastic analysis of a single-server retrial queue with recurrent clients and balking behavior under extended Bernoulli vacations [J]. *Heliyon*, 2024, 10(13): e33919. https://doi.org/10.1016/j.heliyon.2024.e33919

Ke J C, Liu T H, Su S, Zhang Z G. On retrial queue with customer balking and feedback subject to server breakdowns [J]. *Communications in Statistics-Theory and Methods*, 2022, 51(17): 6049-6063. https://doi.org/10.1080/03610926.2020.1852432

Lv S, Yang M, Wen J. The M/M/1 repairable queueing system with two types of server breakdowns and negative customers [J]. *Engineering Letters*, 2024, 32(1): 77.

Madheswari S P, Krishna Kumar B, Suganthi P. Analysis of M/G/1 retrial queues with second optional service and customer balking under two types of Bernoulli vacation schedule [J]. *RAIRO-Operations Research*, 2019, 53(2): 415-443. https://doi.org/10.1051/ro/2017029

Ma Z, Guo S, Wang R. The virtual machines scheduling strategy based on M/M/c queueing model with vacation [J]. *Future Generation Computer Systems*, 2023, 138: 43-51. https://doi.org/10.1016/j.future.2022.08.001

Guo S, Ma Z, Yu X, Chen L. An energy saving strategy based on M/M/c queueing model with preemptive priority and asynchronous working vacation [J]. *International Journal of Autonomous and Adaptive Communications Systems*, 2022, 15(2): 106-125. https://doi.org/10.1504/IJAACS.2022.123462

Melikov A, Aliyeva S, Sztrik J. Retrial queues with unreliable servers and delayed feedback [J]. *Mathematics*, 2021, 9(19): 2415. https://doi.org/10.3390/math9192415

Mehta P, Jain M. ANFIS and metaheuristics for double orbit retrial queue with complete and working vacations [J]. *Applied Soft Computing*, 2024, 167: 112473. https://doi.org/10.1016/j.asoc.2024.112473

Muthusamy S, Devadoss N, Ammar S I. Reliability and optimization measures of retrial queue with different classes of customers under a working vacation schedule [J]. *Discrete Dynamics in Nature and Society*, 2022, 2022: 6806104. https://doi.org/10.1155/2022/6806104

Nazarov A, Sztrik J, Kvach A, Bérczes T. Asymptotic analysis of finite-source M/M/1 retrial queueing system with collisions and server subject to breakdowns and repairs [J]. *Annals of Operations Research*, 2019, 277(2): 213-229. https://doi.org/10.1007/s10479-018-2894-z

Seenivasan M, Senthilkumar R, Subasri K S. M/M/2 heterogeneous queueing system having unreliable server with catastrophes and restoration [J]. *Materials Today: Proceedings*, 2022, 51: 2332-2338. https://doi.org/10.1016/j.matpr.2021.11.567

Shanmugam B, Saravanarajan M C. Unreliable retrial queueing system with working vacation [J]. *AIMS Mathematics*, 2023, 8(10): 24196-24224. https://doi.org/10.3934/math.20231234

Singh S K, Cruz F R, Gomes E S, Banik A D. Classical and Bayesian estimations of performance measures in a single server Markovian queueing system based on arrivals during service times [J]. *Communications in Statistics-Theory and Methods*, 2024, 53(10): 3517-3546. https://doi.org/10.1080/03610926.2022.2155789

Sun K, Liu Y, Li K. Energy harvesting cognitive radio networks with strategic users: A two-class queueing model with retrials [J]. *Computer Communications*, 2023, 199: 98-112. https://doi.org/10.1016/j.comcom.2022.12.017

Tian R, Wu X, He L, Han Y. Strategic analysis of retrial queues with setup times, breakdown and repairs [J]. *Discrete Dynamics in Nature and Society*, 2023, 2023: 4930414. https://doi.org/10.1155/2023/4930414

Vijayashree K V, Pavithra P. Performance analysis and cost optimization of an M/M/1 queue with N-policy and working breakdown [J]. *Methodology and Computing in Applied Probability*, 2025, 27(3): 68. https://doi.org/10.1007/s11009-025-10196-0

Wang Z, Liu L, Zhao Y Q. Equilibrium customer and socially optimal balking strategies in a constant retrial queue with multiple vacations and N-policy [J]. *Journal of Combinatorial Optimization*, 2022, 43(4): 870-908. https://doi.org/10.1007/s10878-021-00814-1

Wang Z, Wang J. Strategic priority-purchasing and joining rules in a retrial queue [J]. *IMA Journal of Management Mathematics*, 2021, 32(2): 161-194. https://doi.org/10.1093/imaman/dpaa010

Wu S, Lan S. Analysis of repairable discrete-time queueing systems with negative customers, disasters, balking customers and interruptible working vacations under Bernoulli schedule [J]. *Mathematics and Computers in Simulation*, 2025, 232: 102-122. https://doi.org/10.1016/j.matcom.2024.12.018

Xu F, Tian R, Shao Q. Optimal pricing strategy in an unreliable M/G/1 retrial queue with Bernoulli preventive maintenance [J]. *RAIRO-Operations Research*, 2023, 57(5): 2639-2657. https://doi.org/10.1051/ro/2023146

Younes S, Idi M, Robbana R. Performability analysis of multi-service call admission control schemes in LTE networks [J]. *International Journal of General Systems*, 2024, 53(2): 215-253. https://doi.org/10.1080/03081079.2023.2270138

Zhang Y. Optimal pricing analysis of computer networks based on a queueing system with retrial mechanism [J]. *IEEE Access*, 2020, 8: 137490-137500. https://doi.org/10.1109/ACCESS.2020.3012612

Zhang Y, Wang J. Quality-speed trade-offs in customer-intensive services with boundedly rational customers and retrials [J]. *Computers & Industrial Engineering*, 2022, 167: 107983. https://doi.org/10.1016/j.cie.2022.107983

Zhou M, Liu L, Chai X, Wang Z. Equilibrium strategies in a constant retrial queue with setup time and the N-policy [J]. *Communications in Statistics-Theory and Methods*, 2020, 49(7): 1695-1711. https://doi.org/10.1080/03610926.2019.1565779

Li Y, Schlather M, Erdfelder E. A queueing model of visual search [J]. *Journal of Mathematical Psychology*, 2023, 115: 102766. https://doi.org/10.1016/j.jmp.2023.102766