Journal of Contemporary Applied Mathematics
V. 16, No 2, 2026, July-December
ISSN 2222-5498, E-ISSN 3006-3183
https://doi.org/10.62476/jcam161.15

Analysis of MAP/PHF/c retrial queueing model with delayed feedback and non-persistence

Natarajan Aishwarya, Agassi Melikov*, Govindan Ayyappan

Abstract. This study focuses on a multi-server retrial queueing model with an infinite orbit capacity that examines customer behaviors, including balking, lack of persistence, and delayed feedback. Customers approach the system following a Markovian arrival process. We consider the possibility that failure may occur during customer service. If a failure occurs, the customer may either leave and retry later, restart the service from the beginning immediately, or continue from the point of failure. Importantly, the server remains operational after a failure. We utilize the phase-type with failures distribution to model this service time. The system's behavior is represented as a multi-dimensional continuous-time Markov chain whose infinitesimal generator is a level-dependent quasi-birth-and-death process. We demonstrate that this Markov chain falls within the category of asymptotically quasi-Toeplitz Markov chains to determine the ergodicity conditions. Further, we calculate steady-state probabilities and system efficiency metrics for this system. We also present numerical experiments that highlight how different parameters affect system performance.

Key Words and Phrases: Markovian arrival process, phase-type with failures distribution, retrial multi-server queue, delayed feedback, non-persistence, infinite orbit

2010 Mathematics Subject Classifications: 60K25, 90B22, 68M20

1. Introduction

Retrial queueing systems play a crucial role in modeling real-world systems like cloud computing and telecommunication networks. When customers (tasks, users, or requests) encounter busy servers or when the buffer is full, they enter a virtual waiting space known as the orbit. From the orbit, each customer continues making attempts at random intervals to access service, resulting in system dynamics that differ from classical queues. The inaugural monograph on retrial

*Corresponding author.

queues by Templeton and Falin [1] provides an in-depth analysis of key methods and findings associated with single-server retrial queueing systems.

Most of the current research in the area of multi-server retrial queueing systems consider a Poisson process for arrivals and an exponential distribution for the service durations. While mathematically tractable, these simplified models usually can not account for the burstiness and correlation of real-life traffic. The Markovian arrival process (MAP) (as detailed in [2, 3]) is a popular alternative for modeling such complicated and dependent arrivals. Similarly, the phase-type (PH) distribution (see [4]) is commonly advocated for modeling service time distributions. For instance, Dudin and Dudina [5] studied a multi-server retrial queueing system with limited buffer in which orders arriving by MAP are served in batches, with batch sizes limited by the minimum of a fixed minimum batch size and the buffer capacity. Service times come from a group size-dependent PH distribution and a classical retrial strategy is used. Chakravarthy [6] investigated a multi-server retrial queueing model with MAP arrivals and PH service and retrial times, and where customers arriving to find all servers busy join an infinite orbit and retry independently. Because of the complexity of the model and the relative lack of previous research, the author uses simulation. The author also showed the effect of assuming exponential retrial times. Recent works have extended this by adopting the batch Markovian arrival process ($BMAP$) to model arrivals (e.g., [7, 8, 9, 10]).

The analysis of unreliable queueing system has focused on server breakdown models, where servers can fail and become unavailable for service for a random recovery period until it is repaired. Several studies (e.g., [11, 12, 13, 14, 15, 16, 17]) have analyzed different scenarios with server breakdown in queueing systems. However, in many applications such as wireless communications or information transmission, failure during service does not necessarily render a server inoperable. Instead, a service may be interrupted by an failure (packet loss or signal errors), without affecting server functionality, and the impacted customer either retries later, starts service over immediately or resumes from the failure point. Recent research has developed phase-type with failures (PHF) distribution for service times, which accounts for service unreliability rather than server breakdowns (described in [18]). Dudin and Dudina [19] examined multi-server retrial queueing systems that incorporate a MAP arrivals and PHF service time distribution. They designed this model to capture what happens when information transmission channels are not reliable.

In retrial queueing theory, the analysis of the Bernoulli feedback mechanism, particularly delayed feedback where served customers probabilistically rejoin the orbit for another service, has been addressed in several notable studies. The following works provide foundational insights. Ayyappan et al. [20, 21] studied

$M/M/1$ retrial queueing systems with loss and delayed Bernoulli feedback for low-priority customers, who may join an infinite orbit or leave after service. The system allows upto k waiting spaces for high-priority and they exit post-service. In [20], they employed a non-pre-emptive priority discipline, where high-priority customers wait for low-priority service completion, whereas in [21], they used a pre-emptive discipline, interrupting low-priority service for immediate high-priority service, with both studies utilizing Neuts' matrix geometric method [22].

A multi-channel queueing system with a Markov Modulated Poisson Process (MMPP) arrival flow and delayed feedback is analyzed by Melikov et al. [23]. Following service, customers have the option to either permanently exit the system or rejoin a orbit for further service with a state-dependent probability. The retrial group consists of repeated calls (r-calls), which either return to the orbit or exit the system with a state-dependent probability if all channels are busy when the r-call arrives. The authors create techniques for calculating this Markov chain's steady-state probabilities and derived key performance measures. The research by Singla and Kaur [24] looks at a two-state feedback and balking retrial queueing system. Customers discovering the server to be occupied balk or enter the retrial orbit. Unsatisfied consumers after service might become feedback customers and find their way back into the orbit. Arrivals follow a Poisson process; service times are exponential. Difference-differential equations produce transient state probabilities.

The multi-dimensional Markov chain's non-uniformity makes studying retrial queues challenging. Since the chain cannot be categorized as a standard quasi-birth-and-death class due to this invariability, the classic results of Neuts [22] cannot be used for analysis. Due to this structural variability, the framework of Asymptotically Quasi-Toeplitz Markov chains (AQTMC) presented in [25] becomes essential, as it accommodated level-dependent dynamics and provides tractable ergodicity conditions. In our study, we apply the enhanced and fast algorithm from [19], specifically tailored for level-dependent AQTMC, to calculate the systems's steady-state distribution, to reduces the computational complexity while maintaining accuracy.

Novelty and contributions of the model:

1. Integration of MAP arrivals, PHF service times, and customer behaviors (balking, non-persistence, delayed feedback) in a unified multi-server retrial queue framework.
2. The system's dynamics is presented as a continuous-time Markov chain (CTMC) with a level-dependent block tridiagonal form. It is further demonstrated that the Markov chain aligns with the category of AQTMC.

3. Conditions for the long-term stability of the model are established. Computation of steady-state probabilities and key performance metrics.
4. Numerical assessment of how failure rates, retrial policies, and customer behaviors impact system performance.

This paper is structured concisely as follows: In Section 2, we explore the motivation for this research. Section 3 then gives a concise mathematical description of the model we are considering. Section 4 outlines the way the system states are modeled using a multi-dimensional CTMC. In Section 5, we derive the infinitesimal generator for this chain. The criteria required for this chain to be ergodic are established in Section 6. The system's key performance metrics are formulated in Section 7. Numerical illustrations in Section 8 demonstrate how system parameters influence performance measures, while Section 9 offers a conclusion to the paper.

2. Motivation

Cloud computing systems, especially those relying on CPU-core-driven systems like private clouds, must balance efficient resource use with reliable task processing under fluctuating demands. In this system, tasks arrive randomly, seeking access to a fixed, identical and non-expandable pool of cloud computing resources (CPU cores). Each task requires a single CPU core for processing. Traditional queueing models assume that the system has the infrastructure to track and maintain a queue for incoming tasks. But if a task is queued, it retains an open connection to the server, which takes both server and task computation resources. Servers can have hundreds of such connections, which slow down the performance. To prevent this resource consumption and performance overhead, this model does not employ a buffer (queue) for tasks. Therefore, if all CPU cores are occupied upon a task's arrival, the task may abandon the system, or it will be placed in a retry queue (orbit). Tasks in the orbit independently retry accessing a core after a random time interval but may leave the system, reflecting non-persistent behavior, after unsuccessful retrial attempts. During task execution, failures (e.g., transient errors (hardware noise), check-pointing failures (corrupted intermediate state)) may occur, prompting tasks to either retry by rejoining the orbit, restart the service from the beginning, or resume from the failure point, ensuring flexible recovery, while the underlying computational core remains fully operational and available for other tasks. Upon completing the service process, tasks either exit the system permanently or rejoin the orbit for a new service attempt (delayed feedback), aligning with the nature of cloud workflows.

3. Mathematical Model Description

We model the inflow of customers to the system with a MAP, a versatile point process offers a framework for modeling correlated arrivals. This process is characterized by two square matrices: D_0 and D_1, each of order m. The transition rates governing the occurrence (non-occurrence) of customer arrivals constitutes the matrix $D_1(D_0)$. The infinitesimal generator is $\tilde{D} = D_0 + D_1$. The retrial queueing systems are composed of c homogeneous servers with no waiting buffer. Accordingly, an arrival who finds a server free immediately commences service. However, if all the servers are occupied, the customer is directed to the orbit with probability b, $0 \leq b \leq 1$ or abandon the system entirely with complementary probability $\bar{b} = 1 - b$.

All tasks in the so-called orbit independently try to re-access the service. When there are k customers in the orbit, the aggregate retrial rate is θ_k. If there are fewer than c occupied servers at the time of the attempt, then the retrial attempt succeeds, enabling the customer to immediately occupy an available server. If no server is available at the time of the retrial, the customer is lost with probability r, $0 \leq r \leq 1$ or rejoins to the orbit with complementary probability $\bar{r} = 1 - r$.

We are framing the service process with the more general PHF distribution which is an extension of the standard PH distribution, originally proposed in [18]. Consider a CTMC, Φ_t, which can occupy any state from the set $\{1, 2, \ldots, M, M+1, M+2\}$. Let $\boldsymbol{\alpha}$ be the probability row vector which determines the initial probabilities for M transient states. A transition rate matrix T, which governs movements between these transient states and a column vector $\mathbf{T_1}$, which specifies the rate of transition from each transient state to a absorbing state $(M+1)$, indicating successful service completion. Additionally, a column vector $\mathbf{T_2} = -T\mathbf{e} - \mathbf{T_1}$ defines the rate of transition to a semi-absorbing state $(M+2)$, representing a service failure. Upon failure, with probability w_1, the service is halted, and the task is redirected to orbit for retrial; with probability w_2, the service restarts from the initial state, with Φ_t reentering a transient state selected from $\boldsymbol{\alpha}$ and with probability $1 - w_1 - w_2$, the service reverts to continuing from the same transient state where the failure occurred, i.e., Φ_t reenters a state it had been in before moving to $M+2$. The PHF distribution's irreducible representation is the set $(\boldsymbol{\alpha}, T, \mathbf{T_1}, w_1, w_2)$. Note that service failures do not affect server availability; the server remains operational and can immediately continue serving the same customer, restart service, or become idle when disrupted customer leaves to the orbit, depending on the recovery policy.

After a customer successfully completes the service process, this customer either exits the system permanently, with probability f, $0 \leq f \leq 1$, or returns to the orbit for another service attempt, with complementary probability $\overline{f} = 1 - f$. Customers requiring re-service attempts are referred to as feedback customers. For simplicity, the orbit does not distinguish between retrial and feedback customers; both are treated identically. Furthermore, the model permits multiple re-service attempts.

Notation convention: Throughout the paper, we use $\bar{x} = 1 - x$ to denote complementary probabilities for $x \in \{b, f, r\}$. Specifically:

- $\bar{b} = 1 - b$: probability of immediate abandonment (balking)
- $\bar{f} = 1 - f$: probability of feedback (return to orbit after service)
- $\bar{r} = 1 - r$: probability of persistence (rejoining orbit after failed retrial)

The process flow is illustrated in Fig. 1.

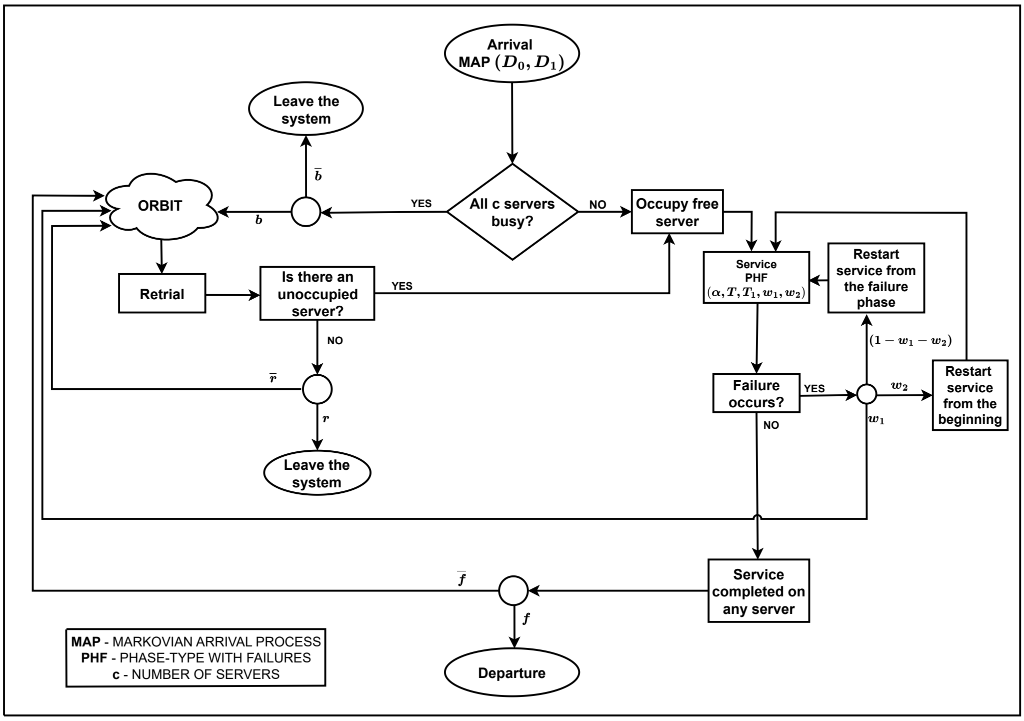
Arrival
MAP (D_0, D_1)
Leave the system
$\bar{b}$
ORBIT
b
YES
All c servers busy?
NO
Occupy free server
Retrial
Is there an unoccupied server?
YES
NO
$\bar{r}$
r
Leave the system
Service PHF $(\alpha, T, T_1, w_1, w_2)$
Restart service from the failure phase
$(1 - w_1 - w_2)$
Failure occurs?
YES
NO
w_2
w_1
Restart service from the beginning
Service completed on any server
$\bar{f}$
f
Departure
MAP - MARKOVIAN ARRIVAL PROCESS
PHF - PHASE-TYPE WITH FAILURES
c - NUMBER OF SERVERS

Figure 1: Schematic representation of the model.

In a multi-server retrial system with n occupied servers and a service time distribution that follows PHF and consists of M transient states, the state space of the multi-dimensional Markov chain that describes the service process on occupied servers is influenced by the tracking method chosen. In the first approach, called track-phase-for-server (TPFS) in [26], we keep track of which service phase each busy server is currently in. The second method, known as count-server-for-phase (CSFP) in [26], counts how many servers are active in each service phase. The CSFP approach results in a significantly smaller state space of size $S_n = \binom{n+M-1}{M-1}$, yielding $n+1 = 11$ states for $M = 2$ and $n = 10$, compared to the TPFS method's state space of size M^n, which produces $2^{10} = 1,024$ states for the same parameters. Due to its smaller state space, the CSFP method offers greater computational efficiency and is therefore preferred for this model.

Now and henceforth, $\mathbf{0}$ indicates a row vector of zeros, O corresponds to a zero matrix, I is an identity matrix, $\mathbf{e}$ denotes a unit column vector, all of corresponding dimensions and $\oplus$ is the Kronecker sum and $\otimes$ is the Kronecker product of matrices.

4. The Process of System States

An irreducible CTMC serves as the mathematical model for the system being analyzed.

$$\tau_t = \{k_t, n_t, \xi_t, \Phi_t^{(1)}, \dots, \Phi_t^{(M)}\}, t \geq 0 \tag{1}$$

where the state variables are defined as follows:

- k_t represents the count of tasks (customers) in the orbit, $k_t \geq 0$;
- n_t represents the count of occupied CPU cores (servers), $n_t = \overline{0, c}$;
- ξ_t represents the arrival phase of MAP, $\xi_t = \overline{1, m}$;
- $\Phi_t^{(i)}$ tracks the count of occupied servers in processing stage i, $i = \overline{1, M}$, $\Phi_t^{(i)} = \overline{0, n_t}$, $\sum_{i=1}^{M} \Phi_t^{(i)} = n_t$.

We use a few specific notations to represent the transition rates of $\Phi_t^{(i)}$, $i \in \{1, 2, \dots, M\}$ as follows:

The matrix $P_n(\boldsymbol{\alpha})$ for $n = \overline{0, c-1}$ defines the transition probabilities of the process $\Phi_t = \{\Phi_t^{(1)}, \dots, \Phi_t^{(M)}\}$, $t \geq 0$, at the instant a new processing task commences, given that n cores are currently occupied.

The matrix $A_n(c, T)$, $n = \overline{0, c}$ specifies the intensities of transitions of Φ_t that do not result in the a jump to $M+1$ or the $M+2$, contingent on n servers being occupied.

For $n = \overline{0, c}$, the matrix $L_{c-n}^{(1)}$, specifies the transition rates of the process Φ_t to the state $M+1$ (absorbing), while $L_{c-n}^{(2)}$ governs the transition rates to the state $M+2$ (semi-absorbing), provided that n cores are occupied. It can be demonstrated that:

$$L^{(1)}_{c-n} = L_{c-n}(c, \tilde{T}_1), \quad L^{(2)}_{c-n} = L_{c-n}(c, \tilde{T}_2),$$

where

$$\tilde{T}_1 = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{T}_1 & O_M \end{pmatrix}, \quad \tilde{T}_2 = \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{T}_2 & O_M \end{pmatrix}.$$

A comprehensive explanation of the matrices $P_n(\boldsymbol{\alpha})$, $A_n(c,T)$, $L^{(1)}_{c-n}$, and $L^{(2)}_{c-n}$ along with the algorithms for the computation of these matrices is given in [27].

Let

$$\Delta_n = -\text{diag}\left\{A_n(c,T)\mathbf{e} + L^{(1)}_{c-n}\mathbf{e} + L^{(2)}_{c-n}\mathbf{e}\right\}, \quad n = \overline{1,c}.$$

The total exit intensities of each state in the process $\Phi_t, t \geq 0$ are given by the modulus of the diagonal entries of the matrix Δ_n, provided that n servers are currently occupied.

5. The Generator Matrix

Having completely described the state space of the process, we now construct the infinitesimal generator $\mathbf{Q}$ of the multi-dimensional CTMC $\{\tau_t, t \geq 0\}$. This infinitesimal generator $\mathbf{Q}$ exhibits a block-tridiagonal structure with level-dependent nature of the orbit dynamics.

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_0^{(0)} & \mathbf{Q}_0^{(1)} & O & O & O & O & \cdots \\ \mathbf{Q}_1^{(-1)} & \mathbf{Q}_1^{(0)} & \mathbf{Q}_1^{(1)} & O & O & O & \cdots \\ O & \mathbf{Q}_2^{(-1)} & \mathbf{Q}_2^{(0)} & \mathbf{Q}_2^{(1)} & O & O & \cdots \\ O & O & \mathbf{Q}_3^{(-1)} & \mathbf{Q}_3^{(0)} & \mathbf{Q}_3^{(1)} & O & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Each block $\mathbf{Q}_k^{(j)}$ is further partitioned according to the number of busy servers $n = 0, 1, \ldots, c$, with sub-blocks $\mathbf{Q}_{n,n'}^{(j)}$ capturing transitions between states with n and n' busy servers. The zero blocks (O) indicate impossible transitions given the model structure.

The blocks are organized as follows:

- $\mathbf{Q}_k^{(0)}$: Transitions within level k (orbit size unchanged). These transitions occur when an event changes the internal phases of customers currently in service or the MAP phase, without altering the orbit size. These include:

  - MAP arrival phase transitions.
  - An arrival occurs when at least one server is free ($n < c$) and immediately starts service—this changes n but not k.
  - An arrival occurs when all servers are busy ($n = c$) and the customer balks (leaves immediately with probability $\bar{b}$)—orbit size unchanged.
  - Phase transitions within the PHF service process.
  - Service failure events where the customer remains in service (with probability $1 - w_1 - w_2$) or restarts service (with probability w_2).
  - A retrial attempt occurs when all servers are busy, and the customer rejoins the orbit with probability $\bar{r} = 1 - r$.
  - Successful service completion where the customer permanently leaves the system with probability f.

- $\mathbf{Q}_k^{(-1)}$: Transitions from level k to $k - 1$ (orbit size decreases). These occur when a customer leaves the orbit, reducing the orbit size by one. Specifically:

  - A retrial attempt succeeds when at least one server is free ($n < c$): the customer leaves the orbit and begins service immediately.
  - When all servers are busy ($n = c$), a retrial attempt fails, and the customer exhibits non-persistence (leaves the system with probability r), decreasing the orbit by one.

- $\mathbf{Q}_k^{(1)}$: Transitions from level k to $k + 1$ (orbit size increases). These occur when a customer joins the orbit, increasing the orbit size by one. Specifically:

  - When a service fails, the customer returns to the orbit with probability w_1.
  - After successful service completion, the customer requires another service with probability $\bar{f} = 1 - f$ and returns to the orbit.
  - When an arrival finds all servers occupied ($n = c$), they join the orbit with probability b.

The exact form of all non-zero sub-blocks is provided in the expressions below;

$$\mathbf{Q}_k^{(0)} = \begin{pmatrix} \mathbf{Q}_{0,0}^{(0)} & \mathbf{Q}_{0,1}^{(0)} & O & \cdots & O & O \\ \mathbf{Q}_{1,0}^{(0)} & \mathbf{Q}_{1,1}^{(0)} & \mathbf{Q}_{1,2}^{(0)} & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & \mathbf{Q}_{c-1,c-1}^{(0)} & \mathbf{Q}_{c-1,c}^{(0)} \\ O & O & O & \cdots & \mathbf{Q}_{c,c-1}^{(0)} & \mathbf{Q}_{c,c}^{(0)} \end{pmatrix},$$

$$\mathbf{Q}_k^{(-1)} = \left(\begin{array}{ccccccc} O & \mathbf{Q}_{0,1}^{(-1)} & O & \cdots & O & O & \\ O & O & \mathbf{Q}_{1,2}^{(-1)} & \cdots & O & O & \\ O & O & O & \cdots & O & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \\ O & O & O & \cdots & \mathbf{Q}_{c-2,c-1}^{(-1)} & O & \\ O & O & O & \cdots & O & \mathbf{Q}_{c-1,c}^{(-1)} & \\ O & O & O & \cdots & O & \mathbf{Q}_{c,c}^{(-1)} & \end{array}\right),$$

$$\mathbf{Q}_k^{(1)} = \begin{pmatrix} O & O & O & \cdots & O & O \\ \mathbf{Q}_{1,0}^{(1)} & O & O & \cdots & O & O \\ O & \mathbf{Q}_{2,1}^{(1)} & O & \cdots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & \cdots & O & O \\ O & O & O & \cdots & O & O \\ O & O & O & \cdots & \mathbf{Q}_{c,c-1}^{(1)} & \mathbf{Q}_{c,c}^{(1)} \end{pmatrix},$$

where

$$\begin{aligned} \mathbf{Q}_{0,0}^{(0)} &= D_0 - \theta_k I_m, \\ \mathbf{Q}_{n,n}^{(0)} &= (D_0 - \theta_k I_m) \oplus (A_n(c,T) + \Delta_n) + (1 - w_1 - w_2) I_m \otimes \mathrm{diag}\{L_{c-n}^{(2)} \mathbf{e}\} \\ &\quad + w_2 I_m \otimes L_{c-n}^{(2)} P_{c-1}(\boldsymbol{\alpha}), \quad n = \overline{1, c-1}, \\ \mathbf{Q}_{c,c}^{(0)} &= (D_0 + \bar{b} D_1) \oplus (A_c(c,T) + \Delta_c) + (1 - w_1 - w_2) I_m \otimes \mathrm{diag}\{L_0^{(2)} \mathbf{e}\} \\ &\quad + w_2 I_m \otimes L_0^{(2)} P_{c-1}(\boldsymbol{\alpha}) - r\theta_k I_{mS_c}, \\ \mathbf{Q}_{n,n+1}^{(0)} &= D_1 \otimes P_n(\boldsymbol{\alpha}), \quad n = \overline{0, c-1}, \\ \mathbf{Q}_{n,n-1}^{(0)} &= f I_m \otimes L_{c-n}^{(1)}, \quad n = \overline{1, c}, \end{aligned}$$

$$\begin{aligned}
\mathbf{Q}^{(-1)}_{n,n+1} &= \theta_k I_m \otimes P_n(\boldsymbol{\alpha}), \quad n = \overline{0, c-1},\\
\mathbf{Q}^{(-1)}_{c,c} &= r\theta_k I_{mS_c},\\
\mathbf{Q}^{(1)}_{n,n-1} &= \overline{f} I_m \otimes L^{(1)}_{c-n} + w_1 I_m \otimes L^{(2)}_{c-n}, \quad n = \overline{1, c},\\
\mathbf{Q}^{(1)}_{c,c} &= bD_1 \otimes I_{S_c}.
\end{aligned}$$

6. Ergodicity

Lemma 1. *The Markov chain* $\tau_t, t \geq 0$, *is classified as continuous-time asymptotically quasi-Toeplitz Markov chains.*

Proof. Following [25], the Markov chain $\tau_t, t \geq 0$, falls within the category of AQTMC if the associated limits exist:

$$Y_0 = \lim_{k\to\infty} R_k^{-1}\mathbf{Q}^{(-1)}_k, \quad Y_1 = \lim_{k\to\infty} R_k^{-1}\mathbf{Q}^{(0)}_k + I, \quad Y_2 = \lim_{k\to\infty} R_k^{-1}\mathbf{Q}^{(1)}_k$$

where R_k is a diagonal matrix formed from the modulus of the corresponding diagonal entries of the matrix $\mathbf{Q}^{(0)}_k$, $k \geq 0$. The problem decomposes naturally into two cases requiring individual consideration:
Case 1: $r = 0$ *(if the customers are persistent)*

$$Y_0 = \begin{pmatrix} O & I_m \otimes P_0(\boldsymbol{\alpha}) & O & \cdots & O\\ O & O & I_m \otimes P_1(\boldsymbol{\alpha}) & \cdots & O\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ O & O & O & \cdots & I_m \otimes P_{c-1}(\boldsymbol{\alpha})\\ O & O & O & \cdots & O \end{pmatrix},$$

$$Y_1 = \begin{pmatrix} O & O & \cdots & O & O\\ O & O & \cdots & O & O\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ O & O & \cdots & O & O\\ O & O & \cdots & U_0 & U_1 \end{pmatrix}, \quad Y_2 = \begin{pmatrix} O & O & \cdots & O & O\\ O & O & \cdots & O & O\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ O & O & \cdots & O & O\\ O & O & \cdots & V_0 & V_1 \end{pmatrix},$$

where

$$\begin{aligned}
U_0 &= W^{-1}\left(f I_m \otimes L^{(1)}_0\right),\\
U_1 &= W^{-1}((D_0 + \overline{b}D_1) \oplus (A_c(c,T) + \Delta_c) + (1 - w_1 - w_2)I_m \otimes \mathrm{diag}\{L^{(2)}_0\mathbf{e}\}\\
&\quad + w_2 I_m \otimes L^{(2)}_0 P_{c-1}(\boldsymbol{\alpha})) + I,\\
V_0 &= W^{-1}\left(\overline{f} I_m \otimes L^{(1)}_0 + w_1 I_m \otimes L^{(2)}_0\right),
\end{aligned}$$

$V_1 = W^{-1}\left(bD_1 \otimes I_{S_c}\right),$

and W is a diagonal matrix formed from the entries of the modulus of the corresponding diagonal entries of the matrix $\mathbf{Q}_{c,c}^{(0)}$ for $k = 0$.
Case 2: $r > 0$ (if the customers are non-persistent)

$$Y_0 = \begin{pmatrix} O & I_m \otimes P_0(\boldsymbol{\alpha}) & O & \cdots & O \\ O & O & I_m \otimes P_1(\boldsymbol{\alpha}) & \cdots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \cdots & I_m \otimes P_{c-1}(\boldsymbol{\alpha}) \\ O & O & O & \cdots & I_{mS_c} \end{pmatrix}, \ Y_1 = O, \ Y_2 = O.$$

In both the cases, Y_0, Y_1 and Y_2 matrices exist and thereby confirms the assertion of Lemma 1.

Theorem 1. *When customers are persistent ($r = 0$), the Markov chain $\tau_t, t \geq 0$, remains ergodic provided the condition $f\mu > b\lambda$ is satisfied, where $\mu = \boldsymbol{\eta} L_0^{(1)}\mathbf{e}$ and $\boldsymbol{\eta}$ is the unique vector solving*

$$\boldsymbol{\eta}\Big[A_c(c,T) + \Delta_c + \left(L_0^{(1)} + (w_1 + w_2)L_0^{(2)}\right)P_{c-1}(\boldsymbol{\alpha}) \\ +(1 - w_1 - w_2)\,\mathrm{diag}\{L_0^{(2)}\mathbf{e}\}\Big] = \mathbf{0}, \qquad (2)$$

$$\boldsymbol{\eta}\mathbf{e} = 1. \qquad (3)$$

For non-persistent customers ($r > 0$), the Markov chain $\tau_t, t \geq 0$, is ergodic regardless of the system parameter settings.

Proof. To derive the ergodicity conditions, we make use of the characteristic that the Markov chain $\tau_t, t \geq 0$ is classified as a continuous-time AQTMC.

As demonstrated in [25], the Markov process $\tau_t, t \geq 0$, possesses a stationary distribution provided the subsequent inequality

$$\boldsymbol{\pi} Y_0 \mathbf{e} > \boldsymbol{\pi} Y_2 \mathbf{e} \qquad (4)$$

is satisfied, where the row vector $\boldsymbol{\pi} = \left(\boldsymbol{\pi}_{(0)}, \boldsymbol{\pi}_{(1)}, \ldots, \boldsymbol{\pi}_{(c)}\right)$ satisfies

$$\boldsymbol{\pi}\left(Y_0 + Y_1 + Y_2\right) = \boldsymbol{\pi}, \quad \boldsymbol{\pi}\mathbf{e} = 1. \qquad (5)$$

We begin by analysing the case for persistent customer (r = 0), by substituting the corresponding Y_0, Y_1 and Y_2 into Eq (5), we get,

$\boldsymbol{\pi}_{(n)} = 0, \quad n = \overline{0, c-2},$

$$\boldsymbol{\pi}_{(c)}\left(U_0+V_0\right)=\boldsymbol{\pi}_{(c-1)}, \tag{6}$$

$$\boldsymbol{\pi}_{(c-1)}\left(I_m\otimes P_{c-1}(\boldsymbol{\alpha})\right)+\boldsymbol{\pi}_{(c)}\left(U_1+V_1\right)=\boldsymbol{\pi}_{(c)}. \tag{7}$$

By solving the above two equations, we get,

$$\boldsymbol{\pi}_{(c)}W^{-1}\Big(I_m\otimes(A_c(c,T)+\Delta_c+(L_0^{(1)}+(w_1+w_2)L_0^{(2)})P_{c-1}(\boldsymbol{\alpha}) \\ +(1-w_1-w_2)\operatorname{diag}\{L_0^{(2)}\mathbf{e}\})+\tilde{D}\otimes I_{S_c}\Big)=0.$$

By substituting directly into the previous equation, it becomes evident that the vector $\boldsymbol{\pi}_{(c)}W^{-1}$ takes the form

$$\boldsymbol{\pi}_{(c)}W^{-1}=a(\boldsymbol{\zeta}\otimes\boldsymbol{\eta})$$

where the normalizing constant, denoted as a, followed by $\boldsymbol{\zeta}$, representing the stationary distribution vector of the MAP process satisfying $\boldsymbol{\zeta}\tilde{D}=\mathbf{0}$ and $\boldsymbol{\zeta}\mathbf{e}=1$, and $\boldsymbol{\eta}$, which corresponds to the invariant distribution vector $\Phi_t=\{\Phi_t^{(1)},\ldots,\Phi_t^{(M)}\}$, which is the unique solution solving

$$\boldsymbol{\eta}\Big[A_c(c,T)+\Delta_c+\Big(L_0^{(1)}+(w_1+w_2)L_0^{(2)}\Big)P_{c-1}(\boldsymbol{\alpha}) \\ +(1-w_1-w_2)\operatorname{diag}\{L_0^{(2)}\mathbf{e}\}\Big]=\mathbf{0},$$

$$\boldsymbol{\eta}\mathbf{e}=1.$$

By substituting the vector $\boldsymbol{\pi}_{(c)}W^{-1}=a(\boldsymbol{\zeta}\otimes\boldsymbol{\eta})$ in Eq (6), we obtain $\boldsymbol{\pi}_{(c-1)}$, which has the following form,

$$\boldsymbol{\pi}_{(c-1)}=a(\boldsymbol{\zeta}\otimes\boldsymbol{\eta})\Big(I_m\otimes(L_0^{(1)}+w_1L_0^{(2)})\Big).$$

Thus, the inequality Eq (4) becomes,

$$a(\boldsymbol{\zeta}\otimes\boldsymbol{\eta})\Big(I_m\otimes(L_0^{(1)}+w_1L_0^{(2)})\Big)(I_m\otimes P_{c-1}(\boldsymbol{\alpha}))\mathbf{e}> \\ a(\boldsymbol{\zeta}\otimes\boldsymbol{\eta})\Big(\overline{f}I_m\otimes L_0^{(1)}+w_1I_m\otimes L_0^{(2)}+bD_1\otimes I_{S_c}\Big)\mathbf{e}.$$

Through algebraic manipulation, we derive the inequality

$$f\mu>b\lambda. \tag{8}$$

This proves the theorem for the persistent case ($r=0$).

The inequality $f\mu>b\lambda$ ensures that the long-run rate at which customers permanently depart the system after service ($f\mu$) exceeds the rate at which new arrivals are redirected to the orbit when all servers are busy ($b\lambda$). Here, $\mu = \boldsymbol{\eta} L_0^{(1)}\mathbf{e}$ represents the total rate of service completion when all servers are busy, λ is the fundamental arrival rate of the MAP, and b is the balking probability under full occupancy. This condition prevents the unlimited accumulation of customers in the orbit and guarantees system stability.

Now, consider for the non-persistent case ($r > 0$), by substituting the corresponding Y_0, Y_1 and Y_2 in Eq (4), we get, $\boldsymbol{\pi} Y_0\mathbf{e} > 0$, which implies $1 > 0$. Thus, for the non-persistent case, the ergodicity condition is always satisfied. This proves the theorem for non-persistent case ($r > 0$). Thus, the theorem is proved.

The ergodic nature of $\{\tau_t, t \geq 0\}$ guarantees that its steady-state probabilities exist and are unique, describing the chain's long-term distribution. Its steady state probabilities are

$$\begin{aligned} &\boldsymbol{z}(k, n, \xi, \Phi^{(1)}, \ \ldots, \Phi^{(M)}) \\ &\qquad = \lim_{t\to\infty} P\{k_t = k, n_t = n, \xi_t = \xi, \Phi_t^{(1)} = \Phi^{(1)}, \ldots, \Phi_t^{(M)} = \Phi^{(M)}\}, \end{aligned}$$

where $k \geq 0$, $n = \overline{0, c}$, $\xi = \overline{1, m}$, $\Phi^{(i)} = \overline{0, n}$, $i = \overline{1, M}$, $\sum_{i=1}^{M} \Phi^{(i)} = n$.

Next, we construct the row vectors $\boldsymbol{z}_{(k,n,\xi)}$ of $\boldsymbol{z}(k, n, \xi, \Phi^{(1)}, \ldots, \Phi^{(M)})$, $k \geq 0$, $n = \overline{0, c}$, $\xi = \overline{1, m}$,

$$\boldsymbol{z}_{(k,n)} = (\boldsymbol{z}_{(k,n,1)}, \ldots, \boldsymbol{z}_{(k,n,m)}), \quad n = \overline{0, c},$$

and the vectors

$$\boldsymbol{z}_k = (\boldsymbol{z}_{(k,0)}, \boldsymbol{z}_{(k,1)}, \ldots, \boldsymbol{z}_{(k,c)}), \quad k \geq 0.$$

It is well established that $\boldsymbol{z}_k, k \geq 0$, satisfies

$$(\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots)\mathbf{Q} = \mathbf{0}, \quad (\boldsymbol{z}_0, \boldsymbol{z}_1, \ldots)\mathbf{e} = 1.$$

To derive the steady-state probability vector $\boldsymbol{z}_k$, $k \geq 0$, we employ Algorithm 3 developed by Dudin and Dudina [19]. This algorithm is specifically designed for deriving the stationary probabilities of level-dependent AQTMC, which precisely model our system's dynamics.

Key features of Algorithm 3: Instead of computing and storing a large number of unnecessary matrices G_k, modeling the transitions of the countable component from level $k + 1$ to k, this algorithm strategically reduces redundant G_k matrices. Then derives the stationary probability vectors $\boldsymbol{z}_k$ through normalized vectors x_k, bypassing the need for larger matrices F_k, significantly lowering memory and computational costs, especially for systems with large generator blocks. For a comprehensive derivation and discussion of Algorithm 3, (see [19] (Section 6.3)).

7. Performance Measures

By utilizing the computed steady-state probability vector $\boldsymbol{z}_k$, $k \geq 0$, we establish relevant performance measures for this system.

- Mean number of occupied servers:

$$c_{busy} = \sum_{k=0}^{\infty} \sum_{n=1}^{c} n \boldsymbol{z}_{(k,n)} \mathbf{e}.$$

- Mean number of tasks in the orbit:

$$E_{orbit} = \sum_{k=1}^{\infty} k \boldsymbol{z}_k \mathbf{e}.$$

- Average system load:

$$E_{system} = E_{orbit} + c_{busy}.$$

- The likelihood that a task, upon arrival, enters service immediately:

$$P_{start} = \frac{1}{\lambda} \sum_{k=0}^{\infty} \sum_{n=0}^{c-1} \boldsymbol{z}_{(k,n)} \left(D_1 \otimes I_{S_n} \right) \mathbf{e}.$$

- The chance that a task is lost due to non-persistent:

$$P_{rlost} = \frac{r}{\lambda} \sum_{k=1}^{\infty} \theta_k \boldsymbol{z}_{(k,c)} \mathbf{e}.$$

- Probability that a task is lost due to balking:

$$P_{blost} = \frac{\overline{b}}{\lambda} \sum_{k=0}^{\infty} \boldsymbol{z}_{(k,c)} \left(D_1 \otimes I_{S_c} \right) \mathbf{e}.$$

- Total probability of customers (tasks) lost:

$$P_{lost} = P_{rlost} + P_{blost}.$$

8. Numerical Results

In this part of the study, we will conduct a number of numerical experiments to study the system's performance characteristics. We utilize MAP and the

PHF service time distribution matrices, as originally described in [19]. While conducting the numerical experiments, the arrivals were defined using the associated MAP matrices:

$$D_0 = \begin{pmatrix} -13.52 & 0 \\ 0 & -0.43(8) \end{pmatrix}, \quad D_1 = \begin{pmatrix} 13.43 & 0.09 \\ 0.2(4) & 0.19(4) \end{pmatrix}.$$

This arrival results in the fundamental rate, λ, is 10, with a correlation coefficient of 0.2 for consecutive inter-arrival times.

The service durations in the system are governed by a PHF distribution, characterized by an Erlang distribution of order 2 for the underlying service process, with a phase transition rate of 2, as defined by:

$$\boldsymbol{\alpha} = (1, 0), \; T = \begin{pmatrix} -2 & 1.9 \\ 0 & -2 \end{pmatrix}, \quad \mathbf{T}_1 = (0, 1.9)^{\mathsf{T}}, \quad \mathbf{T}_2 = (0.1, 0.1)^{\mathsf{T}},$$
$$w_1 = 0.5, \quad w_2 = 0.3.$$

In this structure, service completion occurs only from the second phase, while failures may occur from either phase with equal probability $p = 0.05$.

The remaining parameters are fixed as follows: The rate at which customers in the orbit attempt to re-access the system is given by $\theta_k = k\theta$, where $k \geq 1$, having $\theta = 0.5$. Additionally, $r = 0.2$, $b = 0.8$, $f = 0.6$.

The generator block dimensions are given by: $n_{dim} = m\left(1 + \sum_{n=1}^{c} S_n\right)$.

Illustration 1.

This numerical study illustrates the critical need to account for both arrival process correlations and service failure mechanisms in modeling a system. The base model, referred to as $MAP + PHF$, for evaluating their effect, we examine this model alongside two simplified models.

The first simplified model, denoted $M + PHF$, removes correlation in the arrival process by assuming a stationary Poisson arrival process, defined by matrices $D_0 = -10$ and $D_1 = 10$, which has zero correlation with variation coefficient of one.

The second simplified model, labeled $MAP + PH$, ignores service failures. It utilizes a standard phase-type (PH) distribution - specifically, an Erlang distribution with phase rate 2. In contrast, the PHF distribution in the base model includes a 0.05 probability of failure in each service phase.

The main aim here is to identify how performance metrics calculated based on the simplified models $M + PHF$ and $MAP + PH$ are different from those obtained with the $MAP + PHF$ model with the number of servers c, $c = \overline{(1, 30)}$.

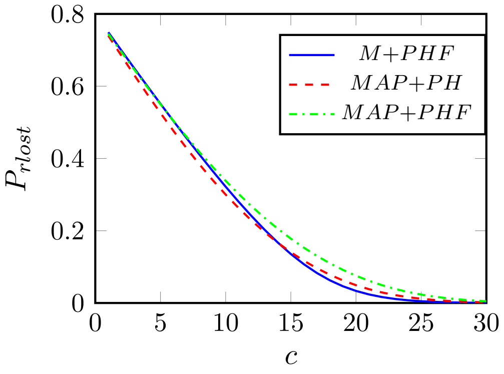

Figure 2: Dependence of loss probability due to non-persistence (P_{rlost}) on the number of servers (c) for three model variants: M+PHF (Poisson arrivals), MAP+PH (reliable service), and MAP+PHF (proposed model).

- As depicted in Figure 2, neglecting either arrival correlation or service failures leads to substantial inaccuracies in performance estimates.
  For instance, at $c = 20$, the loss probability is $P_{rlost} = 0.033438322$ for $M + PHF$, $P_{rlost} = 0.0492359$ for $MAP + PH$, and $P_{rlost} = 0.075504724$ for $MAP + PHF$. This shows that the loss probability of $MAP + PHF$ model is around 1.5 times greater than that of $MAP + PH$ and 2.3 times greater than that of $M + PHF$.
- Analysis of Figures 3 and 4 shows that the approximations $M + PHF$ and $MAP + PH$ drastically underestimate both E_{orbit} and P_{start} as compared to those calculated by the $MAP + PHF$ model.

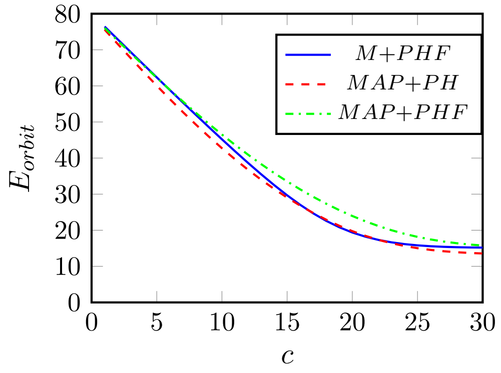

Figure 3: Dependence of the mean number of customers in orbit (E_{orbit}) on the number of servers (c) for three model variants: M+PHF (Poisson arrivals), MAP+PH (reliable service), and MAP+PHF (proposed model).

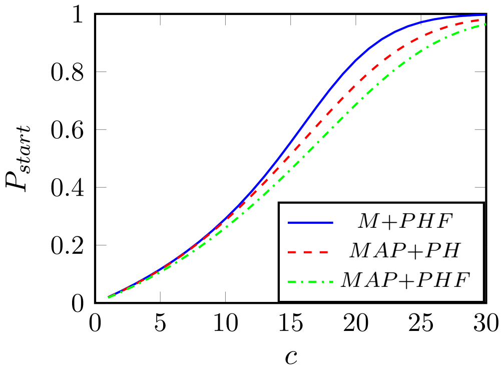

Figure 4: Dependence of the probability of immediate service (P_{start}) on the number of servers (c) for three model variants: M+PHF (Poisson arrivals), MAP+PH (reliable service), and MAP+PHF (proposed model).

These overly optimistic projections indicate that the simplified models do not accurately capture the system's performance characteristics.

Therefore, it is of great importance to consider MAP for arrival process as well as PHF distribution for service times to provide reliable estimation of the system's performance.

Illustration 2.

In prior analyses, the failure probability occurring during service was set to 0.05. In this illustration, we modify this assumption by defining the failure probability at each service phase as p. Consequently, the parameters of the PHF distribution, as outlined in the original model, are adjusted to reflect this variable failure probability.

$$\boldsymbol{\alpha} = (1,0),\ T = \begin{pmatrix} -2 & 2(1-p) \\ 0 & -2 \end{pmatrix},\ \mathbf{T}_1 = (0, 2(1-p))^{\intercal},$$
$$\mathbf{T}_2 = (2p, 2p)^{\intercal},\ w_1 = 0.5,\ w_2 = 0.3.$$

For investigating the effect of varying the failure probability at each service phase, we vary p within the range [0.01, 0.6] and compare various system performance measures.

- As illustrated in Figure 5, the higher the failure probability p gets, the higher the probability of customer loss P_{rlost} becomes. By configuring a maximum acceptable threshold for the loss probability, we can identify the maximum permissible failure probability per service phase. For instance, to meet the requirement $P_{rlost} < 0.1$, failure probability p cannot be more than

0.09, i.e., $p \leq 0.09$. As the failure probability p increases, the likelihood that an arriving customer immediately begins service P_{start} substantially decreases.

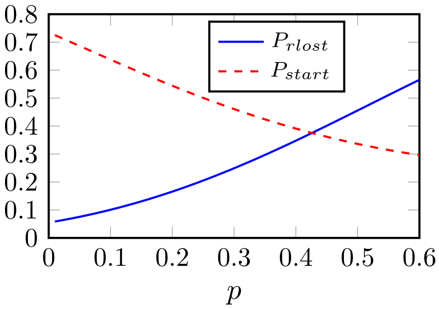

Figure 5: Dependence of loss probability (P_{rlost}) and probability of immediate service (P_{start}) on the failure probability per phase (p).

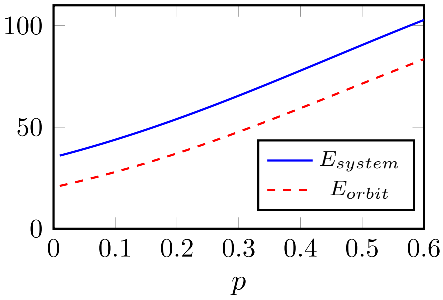

Figure 6: Dependence of mean system size (E_{system}) and mean orbit size (E_{orbit}) on the failure probability per phase (p).

- Analysis presented in Figure 6 demonstrates that both E_{system} and E_{orbit} increase significantly with higher p. This occurs because a rise in p reduces the probability of successful service completion, resulting in more customers entering the orbit after service failure. This leads to higher orbit occupancy, while server utilization may increase due to repeated service after a failure occurs.

Illustration 3.

We examine the influence of retrial rate on the performance metrics. We define the retrial intensities as $\theta_k = k\theta$, where $k \geq 1$, and vary the intensity parameter θ within the range [0.1, 5]. Other system parameters remain consistent with the base model.

- Figure 7 reveals a increase in P_{rlost} as θ rises within the specified interval. This pattern is explained the fact that higher retrial intensities lead to faster retry attempts by customers in the orbit. When all the servers are busy, customers entering the orbit because of server unavailability might try several retries in rapid succession even though the system is still under load. That enhances the chances of customers leaving the system because of non-persistency.
- In Figure 8, E_{orbit} reduces dramatically as θ rises. With increased θ, customers spend fewer moments in the orbit before retry attempts. As a result,

some of these customers successfully go to a server, whereas others are lost due to non-persistence, resulting in fewer customers still in the orbit.

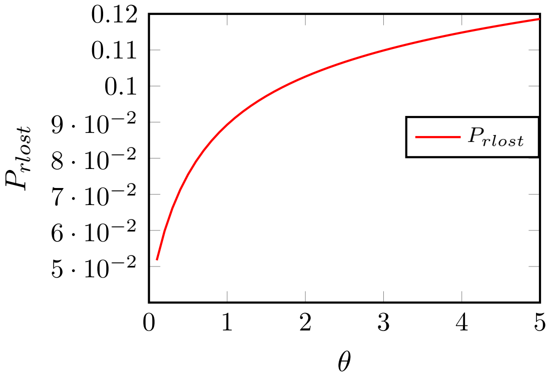

Figure 7: Dependence of loss probability due to non-persistence (P_{rlost}) on the retrial rate parameter (θ).

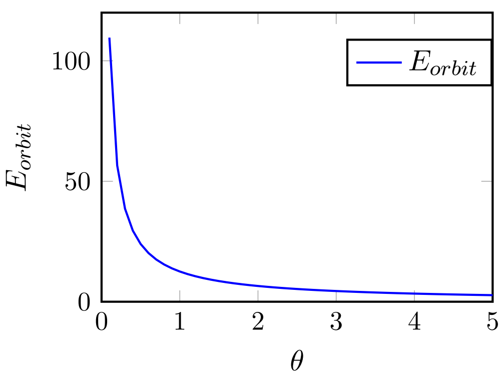

Figure 8: Dependence of mean number of customers in orbit (E_{orbit}) on the retrial rate parameter (θ).

Illustration 4.

We examine the impact of varying the probability r (that a customer abandon the system when all servers are occupied during retry attempts) over the interval [0.1, 0.9].

- The result depicted in Figure 9 shows that as higher probability r increases the likelihood that customers exit the system after failing to access a server, reducing reduces the queue length in both the orbit (E_{orbit}) and the overall system (E_{system}.)

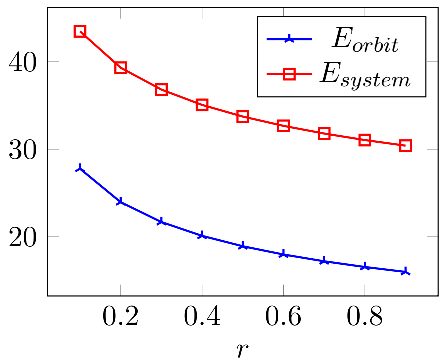

Figure 9: Dependence of mean orbit size (E_{orbit}) and mean system size (E_{system}) on the non-persistence probability (r).

- Observation from Figure 10 indicates that higher r directly increases P_{rlost} because more customers exit the system without service after unsuccessful retrials. The increase in P_{start} occurs because fewer customers in the orbit reduce the overall system load, which increases the likelihood of immediate service for new arrivals.

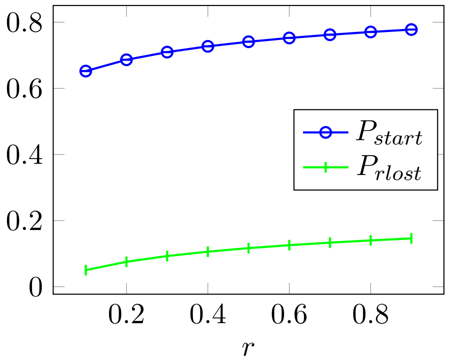

Figure 10: Dependence of probability of immediate service (P_{start}) and loss probability (P_{rlost}) on the non-persistence probability (r).

9. Conclusion

This study investigates a $MAP/PHF/c$ retrial model, incorporating realistic features such as balking, delayed feedback and non-persistence. The system is modeled as a multi-dimensional continuous-time Markov chain, classified as an asymptotically quasi-Toeplitz Markov chains for establishing the ergodic conditions. We calculated the corresponding steady-state probabilities. Numerical experiments and graphical analyses illustrate the effect of system parameters on key performance measures in our model.

To enhance the generality of this model, some extensions would include the addition of batch arrivals and batch service processes to capture bulk data transmissions over communication networks more effectively. Introducing priority-based queueing would enable modeling of differentiated customer classes, while heterogeneous servers that have varying service rates with failures would allow modeling of more realistic distributed systems in telecommunications and cloud computing.

References

[1] Templeton, J.G.C., & Falin, G.I. (1997). Retrial Queues (1st ed.). Routledge.

[2] Chakravarthy, S. R. (2001). The batch Markovian arrival process: A review and future work. Advances in Probability Theory and Stochastic Processes, 21-39.

[3] Chakravarthy, S. R. (2011). Markovian arrival processes. In Wiley encyclopedia of operations research and management science. John Wiley & Sons, Ltd.

[4] Latouche, G., & Ramaswami, V. (1999). Introduction to matrix analytic methods in stochastic modeling. Society for Industrial and Applied Mathematics.

[5] Dudin, A., & Dudina, O. (2024). Retrial queueing system of MAP/PH/N type with a finite buffer and group service. The process describing the system dynamics. In A. Dudin, A. Nazarov, & A. Moiseev (Eds.), Information technologies and mathematical modelling. Queueing theory and applications (pp. 257-271). Springer Nature Switzerland.

[6] Chakravarthy, S. R. (2013). Analysis of MAP/PH/c retrial queue with phase type retrials — Simulation approach. In A. Dudin, V. Klimenok, G. Tsarenkov, & S. Dudin (Eds.), Modern probabilistic methods for analysis of telecommunication networks (pp. 37-49). Springer Berlin Heidelberg.

[7] He, Q. M., Li, H., & Zhao, Y. Q. (2000). Ergodicity of the BMAP/PH/s/s+K retrial queue with PH-retrial times. Queueing Systems, 35, 323-347.

[8] Breuer, L., Dudin, A. & Klimenok, V. (2002). A retrial BMAP/PH/N system. Queueing Systems, 40, 433-457.

[9] Klimenok, V. I., Orlovsky, D. S., & Dudin, A. N. (2007). A BMAP/PH/N system with impatient repeated calls. Asia-Pacific Journal of Operational Research, 24(03), 293-312.

[10] Peng, Y., & Wu, J. (2021). On the $BMAP_1, BMAP_2$/PH/g,c retrial queueing system. Journal of Industrial and Management Optimization, 17(6), 3373-3391.

[11] Chang, F. M., Liu, T. H., & Ke, J. C. (2018). On an unreliable-server retrial queue with customer feedback and impatience. Applied Mathematical Modelling, 55, 171-182.

[12] Chakravarthy, S. R., Shruti, & Kulshrestha, R. (2020). A queueing model with server breakdowns, repairs, vacations, and backup server. Operations Research Perspectives, 7, 100131.

[13] Ke, J. C., Liu, T. H., Su, S., & Zhang, G. (2020). On retrial queue with customer balking and feedback subject to server breakdowns. Communications in Statistics - Theory and Methods, 51(17), 1-17.

[14] Melikov, A., Aliyeva, S., & Sztrik, J. (2021). Retrial queues with unreliable servers and delayed feedback. Mathematics, 9(19), 2415.

[15] Begum, A., & Choudhury, G. (2022). Analysis of an M/$\binom{G_1}{G_2}$/1 queue with Bernoulli vacation and server breakdown. International Journal of Applied and Computational Mathematics, 9.

[16] Jain, M., & Kumar, A. (2022). Unreliable server M/G/1 queue with working vacation and multi-phase repair with delay in verification. International Journal of Applied and Computational Mathematics, 8.

[17] Ismailkhan, E., Narmadha, V., & Yathavan, N. (2025). Analysis of two vacation policies under retrial attempts, Markovian encouraged arrival queueing model. Reliability: Theory & Applications, 20(2(84)), 201-209.

[18] Dudin, A., & Dudin, S. (2016). Analysis of a priority queue with phase-type service and failures. International Journal of Stochastic Analysis, 2016, 1-11.

[19] Dudin, S., & Dudina, O. (2019). Retrial multi-server queuing system with PHF service time distribution as a model of a channel with unreliable transmission of information. Applied Mathematical Modelling, 65, 676-695.

[20] Ayyappan, G., Subramanian, A. M. G. & Sekar, G. (2010). M/M/1 retrial queueing system with loss and feedback under non-pre-emptive priority service by matrix geometric method. Applied Mathematical Sciences, 4(45-48), 2379-2389.

[21] Ayyappan, G., Subramanian, A. M. G., & Sekar, G. (2010). M/M/1 retrial queuing system with loss and feedback under pre-emptive priority service. International Journal of Computer Applications, 2(6), 27-34.

[22] Neuts, M. F. (1981). Matrix-geometric solutions in stochastic models: An algorithmic approach. John Hopkins University Press.

[23] Melikov, A., Aliyeva, S., & Sztrik, J. (2019). Analysis of queueing system MMPP/M/K/K with delayed feedback. Mathematics, 7(11), 1128.

[24] Singla, N., & Kaur, H. (2021). A two-state retrial queueing model with feedback and balking. Reliability: Theory & Applications, 16, 142-155.

[25] Klimenok, V., & Dudin, A. (2006). Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory. Queueing Systems, 54, 245-259.

[26] He, Q. M., & Alfa, A. S. (2018). Space reduction for a class of multidimensional Markov chains: A summary and some applications. INFORMS Journal on Computing, 30(1), 1-10.

[27] Kim, C., Dudin, S., Taramin, O., & Baek, J. (2013). Queueing system MAP/PH/N/N+R with impatient heterogeneous customers as a model of call center. Applied Mathematical Modelling, 37(3), 958-976.

Natarajan Aishwarya
Department of Mathematics, Puducherry Technological University, Puducherry, India
E-mail: 2401714001@ptuniv.edu.in

Agassi Melikov
Department of Mathematics, Baku Engineering University, Khirdalan, Azerbaijan
E-mail: amelikov@beu.edu.az

Govindan Ayyappan
Department of Mathematics, Puducherry Technological University, Puducherry, India
E-mail: ayyappan@ptuniv.edu.in

Received 11 May 2025
Accepted 02 December 2025