RT&A, No 2 (78)
Volume 19, June, 2024

ANALYSIS OF A SINGLE SERVER SYSTEM WITH HETEROGENEOUS ARRIVAL, HETEROGENEOUS SERVICE, SYSTEM FAILURE AND MAINTENANCE

Mohammed Shapique A, Vaithiyanathan A

•

IFET College of Engineering, Villupuram, India
shapique@gmail.com, vaithi05@gmail.com

Abstract

This paper investigates a single-server queuing system with heterogeneous service, failure, and maintenance. The proposed model features a server acting as both the main and backup server. System failure can occur at any stage. When a failure happens, instead of stopping the service entirely, the main server functions as a backup, providing service at a reduced rate. Once all jobs in the system have been serviced, the backup server enters the maintenance state. Following the repair process during maintenance, the server transitions to an idle state, awaiting incoming jobs. Explicit expressions for both transient and steady-state behaviours of the system are derived. Additionally, key system performance metrics are discussed in this paper, accompanied by graphical illustrations to visualize system size probabilities and performance indices.

Keywords: Heterogeneous service; Generating function; Continued fraction; Modified Bessel function, Time-dependent probabilities, Steady-state probabilities

1. Introduction

Queuing systems, fundamental to understanding the dynamics of service provision in various domains, have traditionally been modelled under the assumption of homogeneity, where service rates remain constant across servers. However, the real-world landscape presents a diverse array of scenarios where servers exhibit heterogeneous characteristics, ranging from differing capacities to varied processing speeds. This departure from homogeneity introduces complexities that demand novel modelling approaches to accurately capture system behaviours. In this paper, we delve into the realm of heterogeneous servers within queuing systems, focusing on the intricate interplay between server diversity and system resilience. Our investigation aims to address the challenges posed by system failures, a ubiquitous occurrence in service environments, by proposing a resilient model where servers seamlessly transition between primary and backup roles to ensure continuity of service provision. Specifically, we contribute to the literature by analyzing a single-server queuing system providing two types of service: fast and slow. Instead of halting service entirely during failure, our proposed model allows the server to transition into a backup role and continue providing service at a reduced rate, thus minimizing downtime and enhancing operational resilience.

Several authors have explored queuing systems with heterogeneous servers. For instance, Kumar and Madheswari [8] utilized a Markovian queue model to investigate a system featuring two servers with different characteristics and multiple vacation periods. Using the matrix geometric method, they determined the stationary queue length distribution and average system size for this setup. Krishnamoorthy and Sreenivasan [9] analyzed an M/M/2 queuing system

with two servers of different types. One server remains continuously available, while the other server goes on vacation when no customers are waiting for service. Upon returning from vacation, the second server operates at a reduced rate if the first server is already busy. The authors examined the system's behaviour in a steady state using the matrix geometric method.

Efrosinin and Rykov [5] analyzed a multi-server system with heterogeneous exponential queues. Their study demonstrates techniques for computing steady-state probabilities and deriving distributions for waiting and sojourn times. Efrosinin et al. [6] investigated a controllable multi-server heterogeneous queueing system in which servers operate at different service rates without preemption. Additionally, the authors have applied the concept of heterogeneity in service to cloud centres. Wang et al. [13] introduced the concept of heterogeneous servers in cloud centres to strike an optimal balance between expected response time and power consumption. By incorporating servers with varying capabilities, they aimed to efficiently handle stochastically arriving requests in cloud environments. From the literature survey, it is observed that many authors have focused on utilizing two servers to provide heterogeneous service, with both servers operating at different speeds. However, in this paper, we depart from this convention by considering a single server capable of providing two distinct services. For instance, imagine a modern banking system where a single ATM offers both cash withdrawal and deposit services, catering to the diverse needs of customers. This type of service is also applied in cloud computing. In a cloud computing platform, a single virtual machine instance may be tasked with handling both high-priority real-time data processing and lower-priority batch processing tasks. Additionally, while traditional heterogeneous server models assume a fixed arrival rate, our proposed model introduces heterogeneity in the arrival rate as well, reflecting real-world scenarios where incoming requests vary in frequency and urgency.

In service systems, customers often experience heterogeneous service, which can stem from various reasons. In this paper, we focus on addressing the challenges posed by system failures resulting from technical anomalies, a scenario ubiquitous in real-world service environments. System failures can occur due to several reasons such as negative customers [7], disaster ([3], [11]) and catastrophes [4]. Ammar [2] investigated the two-processor heterogeneous system with catastrophes, server failures and repairs. Sudhesh and Savitha studied three heterogeneous systems with catastrophes. From the literature survey, it is observed that many authors have considered that when a system encounters a disaster, all customers are removed from the system, and the system switches to a failure state. After the repair process, the server switches to an idle state and waits for customers to arrive.

In response to such disruptions, our proposed model incorporates a resilient mechanism wherein the primary server seamlessly transitions into a backup role whenever a failure occurs. During these periods of contingency, the backup server delivers service at a reduced rate, thereby mitigating the impact of disruptions on service provision and maintaining a degree of continuity for system users. Upon serving all customers in the system, the backup server switches to the maintenance state, initiating necessary repairs to restore the system to full functionality. This proactive approach to maintenance ensures the integrity and reliability of the system, minimizing downtime and enhancing overall operational resilience. By integrating these aspects into our queuing model, we aim to provide a comprehensive framework for analyzing and optimizing the performance of service-oriented systems under diverse operating conditions. The objective of this paper is to analyze a single-server queueing system where the server provides two types of service: fast and slow. Instead of halting service entirely during failure, the server transitions into a backup role and continues providing service at a reduced rate. Once all customers have been served, the backup server switches to a maintenance state. Following maintenance, the server returns to an idle state and waits for customers to arrive. To analyze this system, we derive both transient and steady-state probabilities using Laplace transform and generating function techniques.

This article is structured as follows: Section 2 presents the application of the proposed model. Section 3 provides the model description. The time-dependent probabilities of the system are discussed in Section 4, while Section 5 focuses on the performance measures of the system in the

transient state. In Section 6, the steady-state probabilities are presented, followed by a discussion on the performance indices of the system in the steady state in Section 7. A numerical illustration of the system is provided in Section 8, and Section 9 offers the conclusion of the proposed work.

2. Application of the proposed system

The proposed system is applied in Disaster Recovery Systems, which are crucial components of critical IT infrastructure such as data centres or cloud-based services where high availability is essential. A disaster recovery system ensures business continuity and data integrity in the face of unexpected events like hardware failures, natural disasters, or cyber-attacks. In this system, the main server is responsible for handling regular operations and serving client requests. Meanwhile, the backup server operates in a standby mode, continuously replicating data and configurations from the active server to ensure that it remains up-to-date with the latest data.

In the event of a system failure on the main server, the backup server automatically takes over the responsibilities of the main server in a process known as fail-over. This fail-over mechanism may be triggered either manually or automatically by monitoring systems that detect the failure of the main server. Once the main server is repaired and ready to operate again, it can resume its regular duties, and the data changes that occurred during the fail-over period can be synchronized back to the main server. The main server acting as a backup server in this context provides redundancy and enhances the overall reliability of the system. It ensures that critical services and applications remain available even during unexpected disruptions, thereby reducing downtime and minimizing the impact on end-users or customers.

3. Model Description

Consider a system that consists of a single server acting as the main server and also a backup server, providing different types of service. Whenever a failure occurs in the main server, the backup server acts as the main server but with a slower service rate, denoted by μ_2. Arrival occurs to the main server according to a Poisson process with rate λ_1, whereas arrivals occur with rate λ_2 when the backup server is active. Customers receive service at the main server with exponential rate μ_1, while the backup server has a reduced service rate μ_2, where $\mu_2 \leq \mu_1$. Assume that failures of the main server occur at an exponential rate γ. Once the backup server becomes idle, it promptly enters a state of preventive maintenance (state V), characterized by an exponentially distributed duration with a mean of $1/\xi$. Throughout the maintenance period, customers are prohibited from entering the system. The moment the server's maintenance is finished, it promptly transitions back to the primary processor and becomes prepared to attend the new customers.

Let $\{N(t), M(t) : t \geq 0\}$ be the 2-dimensional continuous time Markov chain. Let $\{N(t), t \geq 0\}$ denote the number of customers in the system at any time t and $\{M(t), t \geq 0\}$ represents the state of the system at any time t with state space

$$S = \{(0,0) \cup \{(n,r), n \in Z^+, r = 1,2\} \cup V\}.$$

The state $(0,0)$ represents that the server is idle and waiting for customers to arrive. The state $(n,1)$ represents the main server is busy and providing service to the n^{th} customer. The state $(n,2)$ represents the backup server is busy and providing service to the n^{th} customer. The state V represents the server is in a maintenance state and the server is inoperative in this state. Let $P_{n,r}(t) = P\{N(t) = n, M(t) = r\}$ be the probability that the server is in state r with n number of customers in the system at any time t and let $P_V(t)$ denote the probability that the server is in

maintenance state. Then $P_{n,r}(t)$ and $P_V(t)$ satisfies the following forward Kolmogorov equations

$$P_V^{'}(t) = -\xi P_V(t) + \mu_2 P_{1,2}(t), \tag{1}$$
$$P_{0,0}^{'}(t) = -\lambda_1 P_{0,0}(t) + \xi P_V(t) + \mu_1 P_{1,1}(t), \tag{2}$$
$$P_{1,1}^{'}(t) = -(\lambda_1 + \mu_1 + \gamma) P_{1,1}(t) + \lambda_1 P_{0,0}(t) + \mu_1 P_{2,1}(t), \tag{3}$$
$$P_{n,1}^{'}(t) = -(\lambda_1 + \mu_1 + \gamma) P_{n,1}(t) + \lambda_1 P_{n-1,1}(t) + \mu_1 P_{n+1,1}(t), n \geq 2, \tag{4}$$
$$P_{1,2}^{'}(t) = -(\lambda_2 + \mu_2) P_{1,2}(t) + \mu_2 P_{2,2}(t) + \gamma P_{1,1}(t), \tag{5}$$
$$P_{n,2}^{'}(t) = -(\lambda_2 + \mu_2) P_{n,2}(t) + \lambda_2 P_{n-1,2}(t) + \mu_2 P_{n+1,2}(t) + \gamma P_{n,1}(t), n \geq 2. \tag{6}$$

with the initial condition $P_{0,0}(0) = 1$.

4. Time-dependent probabilities

This section presents the time-dependent probabilities of the system being busy when the main server is active, denoted as $P_{n,1}(t)$, when the backup server is active, denoted as $P_{n,2}(t)$, during maintenance, denoted as $P_V(t)$, and in the idle state, denoted as $P_{0,0}(t)$.

4.1. Evaluation of $P_{n,1}(t)$

This section presents the time-dependent probability of the system being busy when the main server is active. Let $\widehat{P}_{n,r}(s)$ denote the Laplace transform of $P_{n,r}(t)$. Taking Laplace Transform on Equation (4) and rearranging, we get

$$\frac{\widehat{P}_{n,1}(s)}{\widehat{P}_{n-1,1}(s)} = \frac{\lambda_1}{(s + \lambda_1 + \mu_1 + \gamma) - \mu_1 \frac{\widehat{P}_{n+1,1}(s)}{\widehat{P}_{n,1}(s)}}.$$

On simplification, we obtain

$$\widehat{P}_{n,1}(s) = \beta_1 \left[\frac{p_1 - \sqrt{p_1^2 - \alpha_1^2}}{\alpha_1} \right] \widehat{P}_{n-1,1}(s).$$

The above equation recursively yields

$$\widehat{P}_{n,1}(s) = {\beta_1}^{(n-1)} \left[\frac{p_1 - \sqrt{p_1^2 - \alpha_1^2}}{\alpha_1} \right]^{(n-1)} \widehat{P}_{1,1}(s), \quad n \geq 2, \tag{7}$$

where

$$p_1 = s + \lambda_1 + \mu_1 + \gamma, \alpha_1 = 2\sqrt{\lambda_1 \mu_1},\ \beta_1 = \sqrt{\frac{\lambda_1}{\mu_1}}.$$

Taking inverse Laplace transform on Equation (7), we get

$$P_{n,1}(t) = \lambda_1 \beta_1^{n-2} e^{-(\lambda_1 + \mu_1 + \gamma)t} \left[I_{n-2}(\alpha_1(t-u)) - I_n(\alpha_1(t-u)) \right] * P_{1,1}(t), \tag{8}$$

where $I_n(t)$ represents modified Bessel function of first kind of order n. Thus the probability that the main server is busy $P_{n,1}(t)$ is expressed in terms of $P_{1,1}(t)$. The expression for $P_{1,1}(t)$ is presented in Equation (22)

4.2. Evaluation of $P_{n,2}(t)$

To obtain the time-dependent probability of $P_{n,2}(t)$, we define a generating function as follows. Let

$$G(z,t) = \sum_{n=1}^{\infty} P_{n,2}(t) z^n$$

Using Equations (5) and (6), we obtain

$$\frac{\partial}{\partial t} G(z,t) = \left[-(_2+\mu_2) + (_2 z + \frac{\mu_2}{z}) \right] G(z,t) + \gamma \sum_{n=1}^{\infty} P_{n,1}(t) z^n - \mu_2 P_{1,2}(t). \tag{9}$$

Solving Equation (9) yields,

$$\begin{aligned} G(z,t) = \gamma \int_0^t \sum_{n=1}^{\infty} P_{n,1}(u) z^n e^{-(_2+\mu_2)(t-u)} e^{-(_2 z + \frac{\mu_2}{z})(t-u)} du \\ - \mu_2 \int_0^t P_{1,2}(u) e^{-(_2+\mu_2)(t-u)} e^{-(_2 z + \frac{\mu_2}{z})(t-u)} du. \end{aligned} \tag{10}$$

Let

$$\alpha_2 = 2\sqrt{_2 \mu_2},\ \beta_2 = \sqrt{\frac{2}{\mu_2}}.$$

Then

$$e^{-(_2 z + \frac{\mu_2}{z})t} = \sum_{n=-\infty}^{\infty} (\beta_2 z)^n I_n(\alpha_2 t). \tag{11}$$

Using Equation (11) in Equation (10) and equating the coefficient of z^n, we arrive

$$\begin{aligned} P_{n,2}(t) = \gamma \int_0^t \sum_{m=1}^{\infty} P_{m,1}(u) e^{-(_2+\mu_2)(t-u)} \beta_2^{n-m} I_{n-m}(\alpha_2(t-u)) du \\ - \mu_2 \int_0^t P_{1,2}(u) e^{-(_2+\mu_2)(t-u)} \beta_2^n I_n(\alpha_2(t-u)) du. \end{aligned} \tag{12}$$

The above holds for $n \leq -1$ with the left-hand side replaced by zero. Using $I_{-n}(x) = I_n(x)$ for $n \geq 1$

$$\begin{aligned} 0 = \gamma \int_0^t \sum_{m=1}^{\infty} P_{m,1}(u) e^{-(_2+\mu_2)(t-u)} \beta_2^{-n-m} I_{n+m}(\alpha_2(t-u)) du \\ - \mu_2 \int_0^t P_{1,2}(u) e^{-(_2+\mu_2)(t-u)} \beta_2^{-n} I_n(\alpha_2(t-u)) du. \end{aligned} \tag{13}$$

From Equations(12) and (13), we get

$$P_{n,2}(t) = \gamma \int_0^t \sum_{m=1}^{\infty} P_{m,1}(u) e^{-(_2+\mu_2)(t-u)} \beta_2^{n-m} [I_{n-m}(\alpha_2(t-u)) - I_{n+m}(\alpha_2(t-u)) du. \tag{14}$$

4.3. Evaluation of $P_V(t)$ and $P_{0,0}(t)$

This section presents the time-dependent probabilities of the maintenance state and idle state. Taking Laplace transform on Equation (1), we obtain

$$\widehat{P}_V(s) = \frac{\mu_2}{s+\xi} \widehat{P}_{1,2}(s). \tag{15}$$

On inversion, we get

$$P_V(t) = \mu_2 e^{-\xi t} * P_{1,2}(t).$$

Taking Laplace transform on (2), we obtain

$$\widehat{P}_{0,0}(s) = \frac{1}{s+_1}\left[1 + \xi\widehat{P}_V(s) + \mu_1\widehat{P}_{1,1}(s)\right]. \tag{16}$$

On inversion, we have

$$P_{0,0}(t) = e^{-_1 t} * \left[\delta(t) + \xi P_V(t) + \mu_1 P_{1,1}(t)\right]. \tag{17}$$

Setting $n = 1$ in Equation (14) and taking Laplace transform, we get

$$\widehat{P}_{1,2}(s) = \widehat{\Phi}(s)\widehat{P}_{1,1}(s), \tag{18}$$

where

$$\widehat{\Phi}(s) = \frac{\gamma}{2}\sum_{m=1}^{\infty}\beta_1^{m-1}\beta_2^{2-m}\left(\frac{p_1 - \sqrt{p_1^2 - \alpha_1^2}}{\alpha_1}\right)^{m-1}\left(\frac{p_2 - \sqrt{p_2^2 - \alpha_2^2}}{\alpha_2}\right)^{m} \tag{19}$$

and

$$p_2 = s + \lambda_2 + \mu_2.$$

Inverting Equation (19), we get

$$\begin{aligned}\Phi(t) = \gamma\lambda_1 \sum_{m=1}^{\infty}\beta_1^{m-1}\beta_2^{1-m}e^{-(\lambda_1+\mu_1+\gamma)t}\left[I_{m-2}(\alpha_1 t) - I_m(\alpha_1 t)\right] * e^{-(\lambda_2+\mu_2)t} \\ \times \left[I_{m-1}(\alpha_2 t) - I_{m+1}(\alpha_2 t)\right].\end{aligned}$$

Taking Laplace Transform on (3), we get

$$\widehat{P}_{11}(s) = \frac{\lambda_1}{s + \lambda_1 + \mu_1 + \gamma}\widehat{P}_{0,0}(s) + \frac{\mu_1}{s + \lambda_1 + \mu_1 + \gamma}\widehat{P}_{2,1}(s). \tag{20}$$

Setting $n = 2$ in Equation (7) and using Equations (16), (15), (18) in Equation (20), after some algebra, we have

$$\widehat{P}_{1,1}(s) = \lambda_1\sum_{k=0}^{\infty}(\mu_1\beta_1)^k\sum_{r=0}^{k}\left(\frac{\lambda_1\mu_2}{\mu_1\beta_1}\right)^r\binom{k}{r}\frac{1}{(s+\lambda_1)^{r+1}}\left(\frac{p_1 - \sqrt{p_1^2 - \alpha_1^2}}{\alpha_1}\right)^{k-r}\sum_{j=0}^{r}\xi^j\binom{r}{j}\left(\frac{\widehat{\Phi}(s)}{s+\xi}\right)^j. \tag{21}$$

On inversion

$$\begin{aligned}P_{1,1}(t) = \frac{\lambda_1^2}{\beta_1}\sum_{k=0}^{\infty}(\mu_1\beta_1)^k\sum_{r=0}^{k}\left(\frac{\lambda_1\mu_2}{\mu_1\beta_2}\right)^r\binom{k}{r}e^{-\lambda_1 t}\frac{t^r}{r!} * e^{-(\lambda_1+\mu_1+\gamma)t}\left[I_{k-r-1}(\alpha_1 t) - I_{k-r+1}(\alpha_1 t)\right] \\ * \sum_{j=0}^{\infty}\xi^j\binom{r}{j}e^{-\xi t}\frac{t^{j-1}}{(j-1)!} * (\Phi(t))^{*j}.\end{aligned} \tag{22}$$

5. Performance Measures

In this section, the expected system size and variance of the proposed model are presented.

5.1. Expected system size

The expected system size, denoted as $E(N(t))$, is defined as follows.

$$E(N(t)) = \sum_{n=1}^{\infty} n \left(P_{n,1}(t) + P_{n,2}(t)\right)$$

Using Equations (3) − (6), we get

$$\frac{d}{dt}E[N(t)] = \lambda_1 P_{0,0}(t) + (\lambda_1 - \mu_1)\sum_{n=1}^{\infty} P_{n,1}(t) + (\lambda_2 - \mu_2)\sum_{n=1}^{\infty} P_{n,2}(t).$$

Integrating,

$$E(N(t)) = \lambda_1 \int_0^t P_{0,0}(u)du + \sum_{n=1}^{\infty}\left[\int_0^t (\lambda_1 - \mu_1)P_{n,1}(u)du + \int_0^t (\lambda_2 - \mu_2)P_{n,2}(u)du\right].$$

5.2. Variance

The variance of the number of customers at time t is defined as

$$V(N(t)) = E[N^2(t)] - E(N(t))^2$$

where

$$E[N^2(t)] = \sum_{n=1}^{\infty} n^2 \left[P_{n,1}(t) + P_{n,2}(t)\right]$$

Using Equations (3) − (6), we obtain

$$\frac{d}{dt}E[N^2(t)] = \lambda_1 P_{0,0}(t) + \sum_{n=1}^{\infty}\Big[\lambda_1(2n+1)P_{n,1}(t) + \mu_1(1-2n)P_{n,1}(t) + \lambda_2(2n+1)P_{n,2}(t) + \mu_2(1-2n)P_{n,2}(t)\Big].$$

Integrating,

$$E[N^2(t)] = \lambda_1 \int_0^t P_{0,0}(u)du + \sum_{n=1}^{\infty}\Big[\lambda_1(2n+1)\int_0^t P_{n,1}(u)du + \mu_1(1-2n)\int_0^t P_{n,1}(u)du + \lambda_2(2n+1)\int_0^t P_{n,2}(u)du + \mu_2(1-2n)\int_0^t P_{n,2}(u)du\Big].$$

where $P_{n,1}(t), P_{n,2}(t)$ *and* $P_{0,0}(t)$ are given in Equations (18), (14) and (17) respectively.

6. Stationary Analysis

This section presents the steady-state analysis of the proposed model. The steady-state equations of the proposed model are as follows.

$$0 = -\xi\pi_M + \mu_2\pi_{1,2},, \tag{23}$$
$$0 = -\lambda_1\pi_{0,0} + \xi\pi_M + \mu_1\pi_{1,1},, \tag{24}$$
$$0 = -\left(\lambda_1 + \mu_1 + \gamma\right)\pi_{1,1} + \lambda_1\pi_{0,0} + \mu_1\pi_{2,1},, \tag{25}$$
$$0 = -\left(\lambda_1 + \mu_1 + \gamma\right)\pi_{n,1} + \lambda_1\pi_{n-1,1} + \mu_1\pi_{n+1,1}, n = 2,3,4,..., \tag{26}$$
$$,0 = -\left(\lambda_2 + \mu_2\right)\pi_{1,2} + \mu_2\pi_{2,2} + \gamma\pi_{1,1},, \tag{27}$$
$$0 = -\left(\lambda_2 + \mu_2\right)\pi_{n,2} + \lambda_2\pi_{n-1,2} + \mu_2\pi_{n+1,2} + \gamma\pi_{n,1}, n = 2,3,4,...., \tag{28}$$

We define a generating function

$$G_i(z) = \sum_{n=1}^{\infty} \pi_{n,i} z^n, i = 1, 2.$$

Using Equations (25) and (26) and summing for $n = 1, 2, 3, ...$, we get

$$G_1(z) = \frac{z}{(z - z_1)(z - \bar{z}_1)} \{\mu_1 \pi_{1,1} - \lambda_1 \pi_{0,0} z\} \tag{29}$$

where

$$z_1 = \frac{(\lambda_1 + \mu_1 + \gamma) + \sqrt{(\lambda_1 + \mu_1 + \gamma)^2 - 4\lambda_1 \mu_1}}{2\lambda_1},$$

$$\bar{z}_1 = \frac{(\lambda_1 + \mu_1 + \gamma) - \sqrt{(\lambda_1 + \mu_1 + \gamma)^2 - 4\lambda_1 \mu_1}}{2\lambda_1}.$$

It is noted that for $\lambda_1 > 0$, $\mu_1 > 0$, $\gamma > 0$, the roots $z_1 > 1$, $0 < \bar{z}_1 < 1$. Setting $z = \bar{z}_1$ in Equation (29), we obtain

$$G_1(z) = \sum_{n=1}^{\infty} \left(\frac{z}{z_1}\right)^n \lambda_1 \pi_{0,0}$$

Comparing the coefficient of z^n in the above expression, we obtain

$$\pi_{n,1} = \lambda_1 \left(\frac{1}{z_1}\right)^n \pi_{0,0} \tag{30}$$

Similarly, using Equations (27) and (28) and summing for $n = 1, 2, 3, ...$, we get

$$G_2(z) = \frac{z\lambda_2}{(z\lambda_2 - \mu_2)(z - 1)} \{\mu_2 \pi_{1,2} - \gamma G_1(z)\} \tag{31}$$

Setting $z = 1$ in (31), after some algebraic manipulation, we get

$$G_2(z) = \frac{\gamma \lambda_1 \lambda_2 z}{\mu_2 \left(1 - \frac{\lambda_2}{\mu_2}\right)(1 - z)} \left[\sum_{n=1}^{\infty} \left(\frac{1}{z_1}\right)^n - \sum_{n=1}^{\infty} \left(\frac{z}{z_1}\right)^n\right] \pi_{0,0}$$

Using Equation (30) in the above expression and equating the coefficients of z^n on both sides, we get

$$\pi_{n,2} = \gamma \lambda_1 \left\{\sum_{i=1}^{\infty} \left(\frac{1}{z_1}\right)^i \sum_{m=1}^{n} \left(\frac{\lambda_2}{\mu_2}\right)^m - \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} \left(\frac{\lambda_2}{\mu_2}\right)^i \left(\frac{1}{z_1}\right)^j\right\} \pi_{0,0} \tag{32}$$

Setting $n = 1$ in the above result and using it in (23), we obtain

$$\pi_M = \frac{\gamma \lambda_1 \lambda_2}{\xi} \sum_{i=1}^{\infty} \left(\frac{1}{z_1}\right)^i \pi_{0,0}. \tag{33}$$

An explicit expression for $\pi_{0,0}$ can be obtained using the normalisation condition as follows.

$$\pi_M + \pi_{0,0+} \sum_{n=1}^{\infty} \pi_{n,1} + \sum_{n=1}^{\infty} \pi_{n,2} = 1. \tag{34}$$

Using the results (30), (32) and (33) in the above condition, we get

$$\pi_{0,0} = \left[1 + \frac{\gamma \lambda_1 \lambda_2}{\xi} \sum_{i=1}^{\infty} \left(\frac{1}{z_1}\right)^i + \gamma \lambda_1 \sum_{n=1}^{\infty} \left\{\sum_{i=1}^{\infty} \left(\frac{1}{z_1}\right)^i \sum_{m=1}^{n} \left(\frac{\lambda_2}{\mu_2}\right)^m - \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} \left(\frac{\lambda_2}{\mu_2}\right)^i \left(\frac{1}{z_1}\right)^j\right\} + \sum_{n=1}^{\infty} \lambda_1 \left(\frac{1}{z_1}\right)^n\right]^{-1}.$$

7. Performance indices

This section presents the expected system size of the proposed model

7.1. Expected system size

Let $E(N_s)$, $E(N_1)$ and $E(N_2)$ denote the expected number of customers in the system, main server and the backup server respectively.

$$E\left(N_S\right) = E\left(N_1\right) + E\left(N_2\right).$$

Using the result (30) and (31), we get

$$E\left(N_1\right) = \frac{\lambda_1 z_1}{\left(1-z_1\right)^2}\pi_{0,0},$$

$$E\left(N_2\right) = \gamma\lambda_1 \sum_{n=1}^{\infty} n \left\{ \sum_{i=1}^{\infty} \left(\frac{1}{z_1}\right)^i \sum_{m=1}^{n} \left(\frac{\lambda_2}{\mu_2}\right)^m - \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} \left(\frac{\lambda_2}{\mu_2}\right)^i \left(\frac{1}{z_1}\right)^j \right\} \pi_{0,0}.$$

Applying Little's formula, the expected number of customers waiting in the system and the queue is given by

$$E\left(W_s\right) = \frac{1}{\lambda_1} E\left(N_1\right) + \frac{1}{\lambda_2} E\left(N_2\right)$$

$$, E\left(W_q\right) = \sum_{n=1}^{\infty} \left(n-1\right) \pi_{n,1} + \sum_{n=1}^{\infty} \left(n-1\right) \pi_{n,2}.$$

8. Numerical illustration

In this section, we provide a numerical illustration of our proposed model. The parameter values are chosen based on the stability conditions $\frac{\lambda_1}{\mu_1} < 1$ and $\frac{\lambda_2}{\mu_2} < 1$. The parameter values are as follows: $\lambda_1 = 0.6$, $\lambda_2 = 0.5$, $\mu_1 = 1.1$, $\mu_2 = 1$, $\gamma = 0.3$, and $\xi = 0.1$. Figures 1 and 2 depict the behaviour of the main server $P_{1,n}(t)$ and the backup server $P_{2,n}(t)$, respectively. We assumed that the initial condition $P_{0,0}(0) = 1$. As a result, the probability curve of $P_{1,n}(t)$ starts at 1 and gradually decreases until it reaches the steady state. Conversely, all other probability curves for $P_{1,n}(t)$ begin at zero, increase initially, and converge to the steady state. Figures 3 and 4 showcase the expected system size and variance of the system for varying values of the arrival rate λ_1. We observe that as the arrival rate increases, the mean and variance graphs also increase. Figures 5 and 6 show the expected system size and variance for different values of the arrival rate λ_2. Figures 7-10 display the stationary probabilities of the system. Figures 7 and 8 provide insights into the probabilities associated with the main and backup servers, respectively. From the graphs, it is observed that as n increases, the probability curves of $\pi_{n,1}$ and $\pi_{n,2}$ decrease and attain the steady state. Finally, Figures 9 and 10 demonstrate the expected system size in the main and backup servers. We notice that as the arrival rate increases, the expected system size $E(N_i)$, where $i = 1, 2$, for both the main and backup servers also increases. This provides important insights into the system's performance under different workload scenarios.

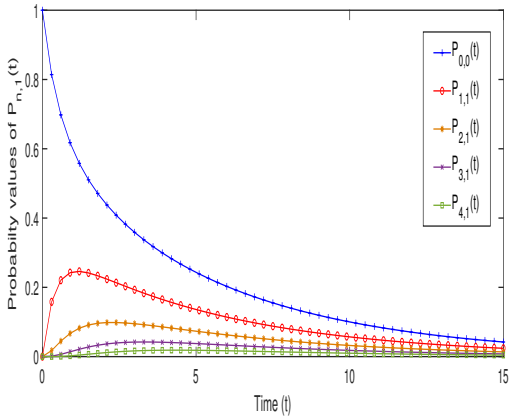

Figure 1: *Probabilities of the main server* $P_{1,n}(t)$.

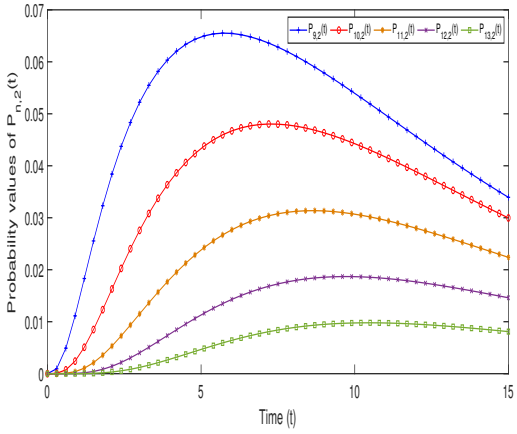

Figure 2: *Probabilities of the backup server* $P_{2,n}(t)$.

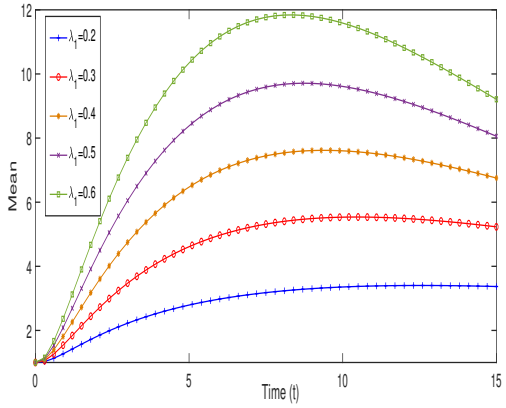

Figure 3: *Mean system size for different arrival rate* λ_1.

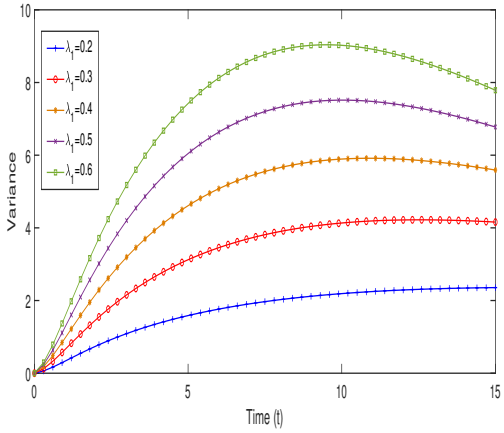

Figure 4: *Variance of the system for different* λ_1.

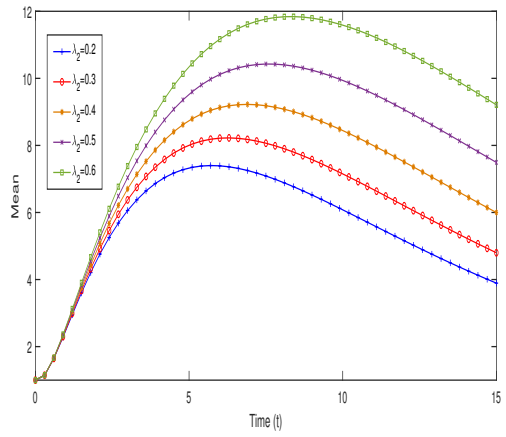

Figure 5: *Mean system size for different* λ_2.

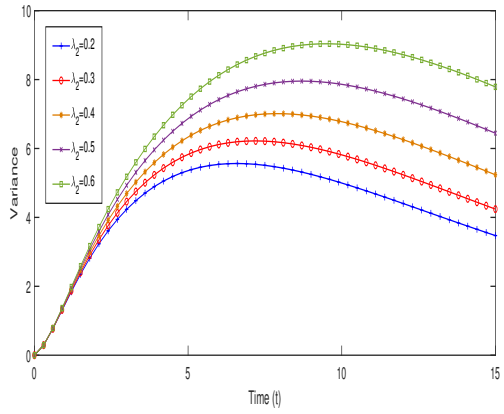

Figure 6: *Variance of the system for different* λ_2.

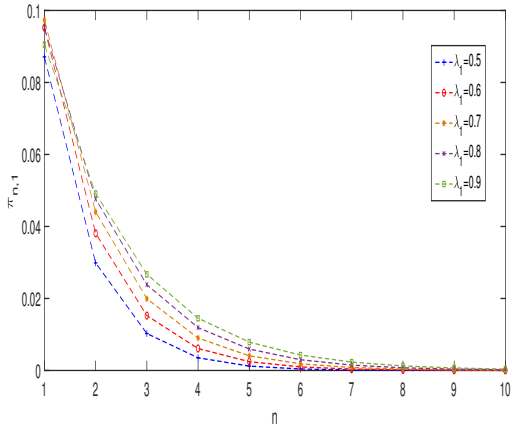

Figure 7: *Steady state probability $\pi_{n,1}$ for different arrival rate λ_1.*

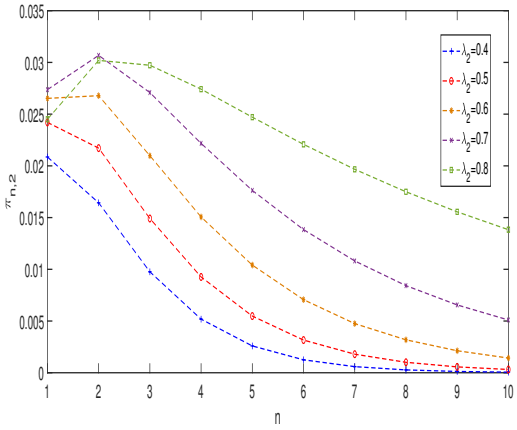

Figure 8: *Steady state probability $\pi_{n,2}$ for different arrival rate λ_2.*

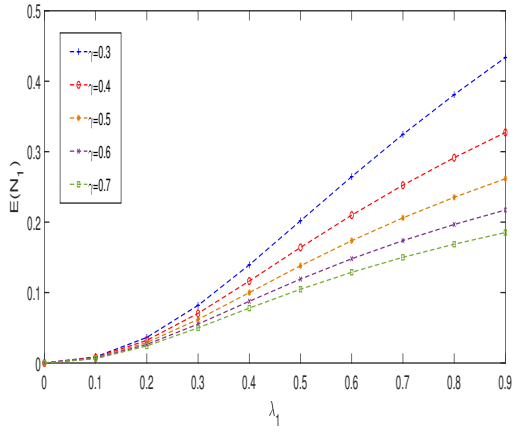

Figure 9: *Mean system size $E(N_1)$ against λ_1 for various γ rates*

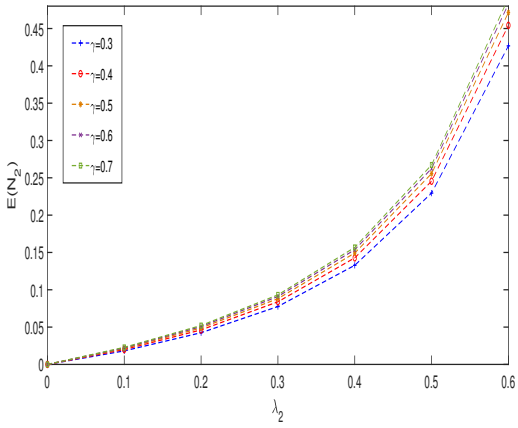

Figure 10: *Mean system size $E(N_2)$ against λ_2 for various γ rates.*

9. Conclusion

This paper investigates an M/M/1 queueing system with heterogeneous service rates and periodic server maintenance. By deriving explicit expressions for both the transient and steady-state probabilities, the study provided a comprehensive understanding of the system's performance under various operating conditions. The establishment of the mathematical framework and the utilization of analytical techniques were instrumental in achieving the desired analysis. The current study focused on a single server setup. One can extend this work by investigating multi-server configurations

References

[1] Anisimov, V., Artalejo, R (2001). Analysis of Markov multi-server retrial queues with negative arrivals, Queueing Systems, 39:157 - 182.

[2] Ammer, I (2014). Transient behaviour of a two-processor heterogeneous system with catastrophes, server failures and repairs, Applied Mathematical Modelling, 38: 2224 - 2234.

[3] Boxma, O. J., Perry, D., & Stadje, W. (2001). Clearing models for M/G/1 queues. Queueing Systems, 38: 287-306.

[4] Di Crescenzo, A., Giorno, V., Nobile, A. G., & Ricciardi, L. M. (2003). On the M/M/1 queue with catastrophes and its continuous approximation. Queueing Systems, 43: 329-347.
[5] Efrosinin, D. V., & Rykov, V. V. E. (2008). On performance characteristics for queueing systems with heterogeneous servers. Automation and Remote Control, 69(1): 61-75.
[6] Efrosinin, D., Stepanova, N., Sztrik, J., & Plank, A. (2020). Approximations in performance analysis of a controllable queueing system with heterogeneous servers. Mathematics, 8(10): 1803.
[7] Gelenbe, E. (1991). Product-form queueing networks with negative and positive customers. Journal of Applied Probability, 28(3): 656-663.
[8] Kumar, B. K., Madheswari, S. P (2005) An M/M/2 queueing system with heterogeneous servers and multiple vacations. Mathematical and Computer Modelling, 41(13): 1415-1429.
[9] Krishnamoorthy, A., Sreenivasan, C (2012). An M/M/2 queueing system with heterogeneous servers including one with working vacation. International Journal of Stochastic Analysis, 2012: 1-16.
[10] Sudhesh, R., Mohammed Shapique, A., Dharmaraja, S (2022). Analysis of a Multiple Dual'Stage Vacation Queueing System with Disaster and Repairable Server. Methodology and Computing in Applied Probability, 24(4): 2485-2508.
[11] Sudhesh, R., Vaithiyanathan, A (2019). Analysis of state-dependent discrete-time queue with system disaster. RAIRO-Operations Research, 53(5): 1915-1927.
[12] Sudhesh, R., & Savitha, P. (2017). Transient behaviour of three-heterogeneous servers queue with system disaster and server repair. RAIRO-Operations Research, 51(4): 965-983.
[13] Wang, S., Li, X., & Ruiz, R. (2019). Performance analysis for heterogeneous cloud servers using queueing theory. IEEE Transactions on Computers, 69(4): 563-576.