

Performance Analysis of N-Policy Finite Queue with 2-Stage Services and Controlled Repairs

Aditya Pratap Singh and Amita Bhagat

Jaypee Institute of Information Technology, Noida, India

Email: adityapratap44@gmail.com, bhagatamita03@gmail.com

(Received April 27, 2023, Accepted September 16, 2023)

Abstract: The paper deals with single server finite capacity queuing model in which arrival of the customers is managed using N-policy. The service is provided to the customers in compulsory 2-stages, wherein after acquiring the service at first stage, the customer move to another stage to complete the service process. The server is unreliable and may fail while providing services. The broken down server is sent to the repair station following the threshold recovery policy. The study also makes use of N-policy, the aforementioned control policy regulates the start of services and also helps for better utilization of the server. The transient analysis of the model has been carried out employing the Runge Kutta method of fourth-order. Various queuing and reliability performance measures are achieved in the study. The cost analysis and sensitivity analysis of the system have been done to explore the effect of different parameters on various performance indices.

Keywords: Unreliable server, N - policy, Threshold recovery, 2-stage services, cost function.

1. Introduction

The queue has become part of our lives in almost all aspects, for example, hospitals, railway stations, shopping complexes, etc. Queues are everywhere, hence queue management is essential for providing a hassle-free pleasing environment to the patients in the hospital, for providing railway tickets to the passengers conveniently, and ensuring the proper space allocation for them. The management of the server is a key problem for queueing systems. Managing the server ensures timely and efficient services adding to the economic quality of the server too by regulating the cost. According to N-Policy, the server starts providing service after the accumulation of N or more customers in the system. As for an example, a new course at any University says that the course will run only if a certain number say N students have

enrolled in it so that course can be sufficient enough to provide funds from admission fees for any associated activities or labs, better engagement of students with each other and better time utility of the mentor providing knowledge. This concept of queue management is very efficient as it also indirectly keeps a check on the quality of services introduced. Yadin and Naor¹ were the first to examine the idea of N-policy in order to look at single server queueing systems. Extension of work by Yadin and Naor¹ was done by Heyman² and they considered M/G/1 queueing system under N-policy and for the optimal value of N, a closed form expression is derived and used a particular cost structure. Wang and Ke³ also studied N-policy queueing systems with fixed capacity.

It is always a matter of concern when services are interrupted due to server breakdown. The server being studied in this investigation is unreliable and might malfunction while offering customer services. As repairing a broken-down server or any part of it demands time as well as the cost is incurred on the system. So, sometimes more efficient queue policies like threshold recovery policy is adopted to repair the broken down server. According to the threshold recovery policy, the remediation process of the server follows a threshold recovery, so remediation will only start when a sufficient number of customers, say q (threshold value), have accumulated in the system. The advantage of this policy lies in the fact by optimizing the expenses that are laid on calling the repairman after threshold accumulation versus after every breakdown. Efrosinin and Semenova⁴ studied an unreliable finite-capacity queueing system in which the repair process is followed by threshold recovery. In addition, Jain *et al.*⁵ used the matrix method to determine state probabilities for queueing systems in which the server is unreliable and follows a threshold recovery policy. Jain and Bhagat⁶ studied a retrial queue model with a defined population using threshold recovery and geometric arrivals. Kumar and Jain⁷, Ayyappan *et al.*⁸, and Ahuja *et al.*⁹ investigated queueing models with unreliable servers in different domains considering a variety of situations. Meena *et al.*¹⁰ worked on a queueing model where a comparative cost analysis using particle swarm optimization and genetic algorithms done for a N-policy M/G/1 queueing system. Chakravarthy *et al.*¹¹ worked on queueing models wherein a backup server is used in the absence of the main server due to breakdown and steady-state analysis have been carried out with various results.

In the present work, authors studied queueing model with 2-stage services and controlled rates. The service mechanism starts only after following N-policy and repair process is initiated only after condition for threshold recovery is satisfied. The concerned model's transient analysis is conducted

with various types of parameters. The time dependent analysis has been done with the help of numerical procedures and many performance measures have been obtained. Section 2 deals with the detailed description of the model including various assumptions, application to the concerned model and governing equations. Sections 3 and 4 respectively, deal with the various performance measures and sensitivity analysis of these measures. The cost analysis is completed in part 5, and section 6 deals with the conclusions and the future utility of the model.

2. Model Description

The model considers a finite-capacity Markovian queueing system with a single server. The model utilizes N -policy, 2-stage services and unreliable server which may stop working any time while providing services. The assumptions used in formulating the model are:

- Arrival of the customers in the Poisson pattern with rate λ . Of course, the inter-arrival time follows exponential distributions.
- Customers keep accumulating in the system in the wait of service till they reach N in the count. The service process initiates as soon as the N -policy is achieved. The server prepares itself and its surroundings to provide service that is done with startup rate ψ .
- The service is provided in 2-compulsory stages wherein after completing the service from stage 1, then customer can avail the services of stage 2. The service rate for stage 1 is denoted by μ_1 and for stage 2 is denoted by μ_2 . Both the service rates follow exponential distributions.
- As realistic, the server may crash down anytime while providing services with rate α_1 while being in stage-1 service and with rate α_2 while being in stage-2, following exponential distribution.
- Broken down server is sent to the repair-facility center for repair and the repair starts following the threshold recovery policy. It means that the repair process starts only if the accumulated customers are q (say) in count at the time of breakdown or the system will wait for the accumulation of q customers to start the repair. This is done to make the system more economical.
- The repair rate follows exponential distribution with rate β . Inter-repair times and inter-failure times follow exponential distributions.

2.1 State of the Server: Let $S(t)$ denote the state of the server at any time t . The state of the server can be defined as:

$$S(t) = \begin{cases} 0; & \text{Idle state (under accumulation of } N \text{ customers)} \\ 1; & \text{Busy at first stage service} \\ 2; & \text{Busy at second stage service} \\ 3; & \text{Repair state} \end{cases}$$

Transient state probabilities are obtained and are denoted by $P_{j,n}(t)$, where the state of the server is denoted by j and strength of the customers is denoted by n in the system at time t .

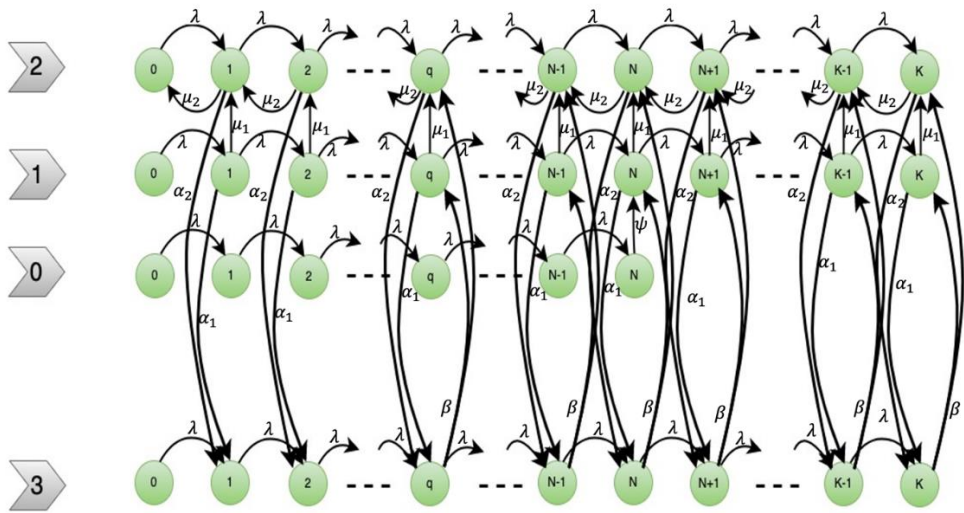


Figure 1: State transition rate diagram for finite capacity queuing system

2.2 Application of the proposed model

Unreliable queuing models under consideration holds a significant area of research while dealing with the real-world congestion and waiting situations. Practically, there are various situations where services start after accumulation of fixed number of customers using the system. Many corporations and organisations like sugar mills, oil refineries etc., starts functioning when a certain amount of resource (N-policy) is available to be processed i.e., they require for some threshold quantity (say N) to be available to start the services. Completion of required service involves multiple subservices to be done in various compulsory stages. The machines processing the resources are unreliable and may breakdown at any time while providing services. Usually, in many situations repair man is called upon for repair if a sufficient number of customers or sufficient amount of material is left for processing (threshold repair policy) owing to minimize the cost of visit of repairman. Queueing models hold their own special position in evaluating the performance of the models and in thus optimizing the cost

corresponding to the system. In addition, the performance analysis also helps in defining the possible ranges in which certain parameters can work better. Hence the proposed model queueing model can be taken as an analogue to the various real-life situations.

2.3 Governing Equations

Here, in this section we define the transient equations for different states of the system. The mathematical equations corresponding to various states of the system are given below:

Server Status $\mathbf{j} = 0, n \geq 0$; idle state of the system

$$(1) \quad \frac{dP_{0,0}(t)}{dt} = -\lambda P_{0,0}; n = 0$$

$$(2) \quad \frac{dP_{0,n}(t)}{dt} = \lambda P_{0,n-1} - \lambda P_{0,n}; 1 \leq n \leq N - 1$$

$$(3) \quad \frac{dP_{0,N}(t)}{dt} = \lambda P_{0,N-1} - \psi P_{0,N}; n = N$$

Server Status $\mathbf{j} = 1, n \geq 1$; busy state (Service-1) of the system

$$(4) \quad \frac{dP_{1,0}(t)}{dt} = -\lambda P_{1,0}; n = 0$$

$$(5) \quad \frac{dP_{1,n}(t)}{dt} = \lambda P_{1,n-1} - (\alpha_1 + \mu_1 + \lambda) P_{1,n}; 1 \leq n \leq q - 1$$

$$(6) \quad \frac{dP_{1,n}(t)}{dt} = \lambda P_{1,n-1} + \beta P_{3,n} - (\alpha_1 + \mu_1 + \lambda) P_{1,n}; q \leq n \leq N - 1$$

$$(7) \quad \frac{dP_{1,n}(t)}{dt} = \lambda P_{1,n-1} + \beta P_{3,n} + \psi P_{0,N} - (\alpha_1 + \mu_1 + \lambda) P_{1,n}; n = N$$

$$(8) \quad \frac{dP_{1,n}(t)}{dt} = \lambda P_{1,n-1} + \beta P_{3,n} - (\alpha_1 + \mu_1 + \lambda) P_{1,n}; N + 1 \leq n \leq K - 1$$

$$(9) \quad \frac{dP_{1,n}(t)}{dt} = \lambda P_{1,n-1} + \beta P_{3,n} - (\alpha_1 + \mu_1) P_{1,n}; n = K$$

Server Status $\mathbf{j} = 2, n \geq 0$; busy state (Service-2) of the system

$$(10) \quad \frac{dP_{2,0}(t)}{dt} = \mu_2 P_{2,1} - \lambda P_{2,0}; n = 0$$

$$(11) \quad \frac{dP_{2,n}(t)}{dt} = \lambda P_{2,n-1} + \mu_1 P_{1,n} + \mu_2 P_{2,n+1} - (\alpha_2 + \mu_2 + \lambda) P_{2,n}; 1 \leq n \leq q - 1$$

$$(12) \quad \frac{dP_{2,n}(t)}{dt} = \lambda P_{2,n-1} + \mu_1 P_{1,n} + \mu_2 P_{2,n+1} + \beta P_{3,n} - (\alpha_2 + \mu_2 + \lambda) P_{2,n}; q \leq n \leq K - 1$$

$$(13) \quad \frac{dP_{2,n}(t)}{dt} = \lambda P_{2,n-1} + \mu_1 P_{1,n} + \beta P_{3,n} - (\alpha_2 + \mu_2) P_{2,n}; n = K$$

Server Status $\mathbf{j} = 3, n \geq 1$; repair state of the system

$$(14) \quad \frac{dP_{3,0}(t)}{dt} = -\lambda P_{3,0}; n = 0$$

$$(15) \quad \frac{dP_{3,n}(t)}{dt} = \lambda P_{3,n-1} + \alpha_1 P_{1,n} + \alpha_2 P_{2,n} - \lambda P_{3,n}; 1 \leq n \leq q - 1$$

$$(16) \quad \frac{dP_{3,n}(t)}{dt} = \lambda P_{3,n-1} + \alpha_1 P_{1,n} + \alpha_2 P_{2,n} - (\lambda + 2\beta) P_{3,n}; q \leq n \leq K - 1$$

$$(17) \quad \frac{dP_{3,n}(t)}{dt} = \lambda P_{3,n-1} + \alpha_1 P_{1,n} + \alpha_2 P_{2,n} - 2\beta P_{3,n}; n = K$$

It is quite a cumbersome task to solve above set of transient equations analytically. So, these equations are solved numerically in MATLAB software using ode45 function solver for first order differential equations. The solver ode45 works on the mathematics of Runge Kutta method of order four.

3. Performance Measures

For the justification of any queueing model, performance measures play a major role. With the help of performance measures, one can decide the practical utility of the model. Some performance measures are given below:

The chances of the server being in idle state at time t

$$(18) \quad P_I(t) = \sum_{n=0}^N P_{0,n}(t)$$

The chances of the server in busy state at time t

$$(19) \quad P_B(t) = \sum_{n=1}^K P_{1,n}(t) + P_{2,n}(t)$$

The chances of the server being in repair state at time t

$$(20) \quad P_R(t) = \sum_{n=q}^K P_{3,n}(t)$$

Expected queue length of the system at time t

$$(21) \quad L_S(t) = \sum_{n=1}^N n P_{0,n}(t) + \sum_{i=1}^3 \sum_{n=1}^K n P_{i,n}(t)$$

Expected waiting time

$$(22) \quad E_w(t) = \frac{L_S(t)}{\lambda}$$

System throughput

$$(23) \quad TP(t) = \mu_1 \sum_{n=1}^K P_{1,n}(t) + \mu_2 \sum_{n=0}^K P_{2,n}(t)$$

Failure frequency

$$(24) \quad Ff(t) = \alpha_1 \sum_{n=1}^K P_{1,n}(t) + \alpha_2 \sum_{n=0}^K P_{2,n}(t)$$

4. Sensitivity analysis

In this section, sensitivity analysis have been done for the validation of proposed model. For the numerical study of various performance measures, MATLAB code using ode45 function is implemented to solve transient

equations. For the calculation of performance measure (18-24), following default values of the parameters have been taken:

$$\lambda = 2, \mu_1 = 7, \mu_2 = 7, \alpha_1 = 0.5, \alpha_2 = 0.3, \beta = 2, \psi = 0.8, N = 8, q = 4, K = 30.$$

4.1 Effect on Expected Waiting Time

Waiting in queues is the most irritating for the customers, thus waiting time ($E_w(t)$) is one of the important performance measures that needs to be minimized for better functioning of the queueing system. Figures 2-4 illustrate the effect of service rate and arrival rate on waiting time. As visible, waiting time decreases with an increase in the service rate whereas it increases for increasing values of arrival rate. Thus, an increase in service rates can help in establishing less crowded system by minimizing waiting time.

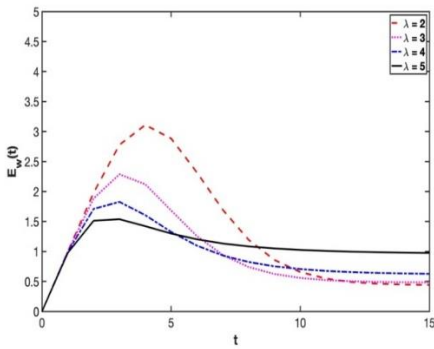


Figure 2: $E_w(t)$ vs λ

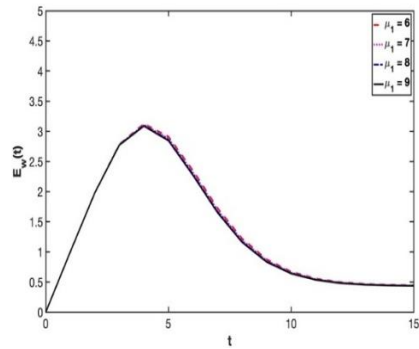


Figure 3: $E_w(t)$ vs μ_1

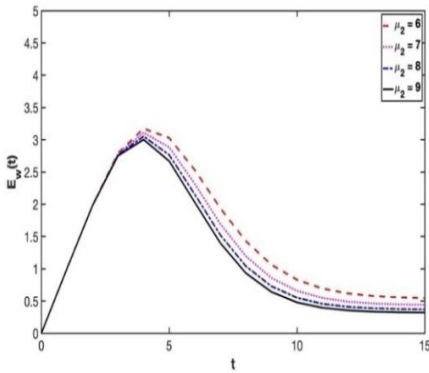


Figure 4: $E_w(t)$ vs μ_2

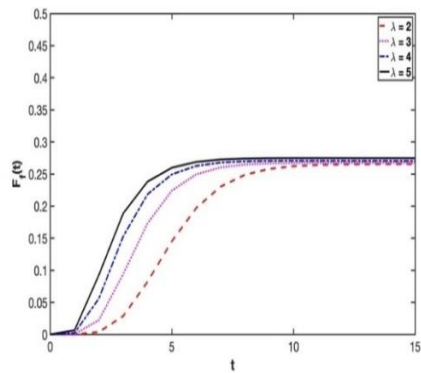


Figure 5: $F_f(t)$ vs λ

4.2 Effect on failure frequency

Effect of arrival rate and service rate on failure frequency ($F_f(t)$) is represented in figures 5-7, which shows that failure frequency of the system increases if either service rate or arrival rate increases. Figure 8, validates the fact that $F_f(t)$ of the system is high for higher values of start-up rate.

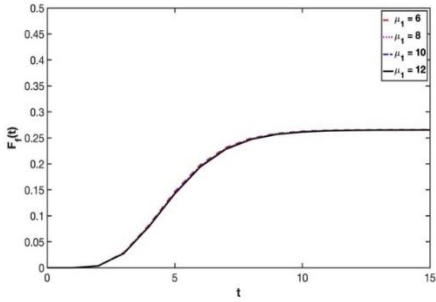


Figure 6: $F_f(t)$ vs μ_1

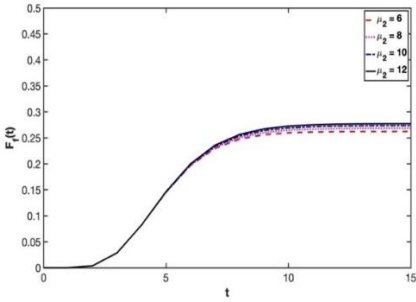


Figure 7: $F_f(t)$ vs μ_2

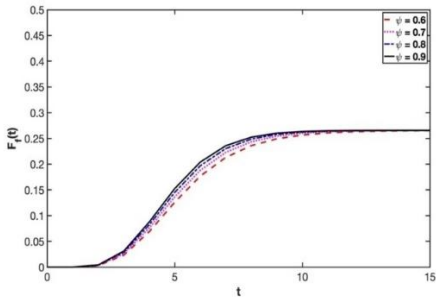


Figure 8: $F_f(t)$ vs ψ

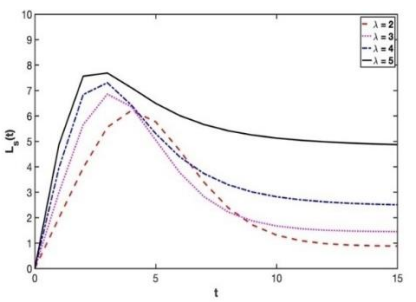


Figure 9: $L_s(t)$ vs λ

4.3 Effect on queue length

Changes in queue length ($L_s(t)$) with varying values of various parameters can be observed in figures 9-14. It is obvious and validated from the figures too that the $L_s(t)$ increases with λ (figure 9) and decreases if the service rates increase (figures 10-11). The server initiates and prepares the surroundings for services with start-up rate ψ , if N-policy is achieved. For larger values of queue length is large. Effect of start-up rate ψ can be seen in figure 12. Variations in queue length with varying values of breakdown rate is depicted in figure 13 and it can be verified that queue length grow if breakdown rate increases. Effect of repair rate on queue length is represented in figure 14.

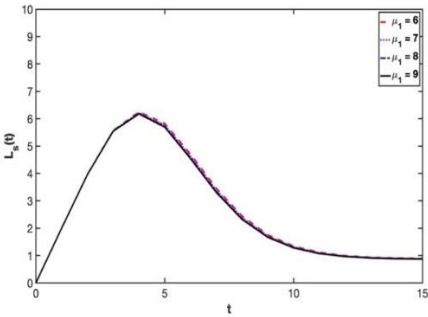


Figure 10: $L_s(t)$ vs μ_1

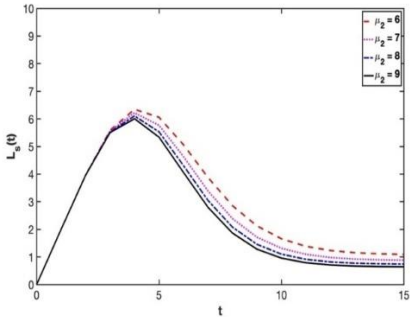
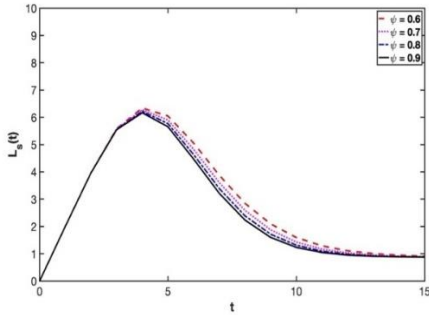
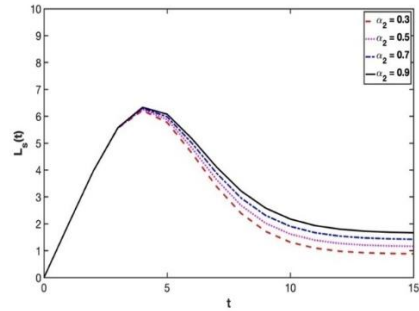
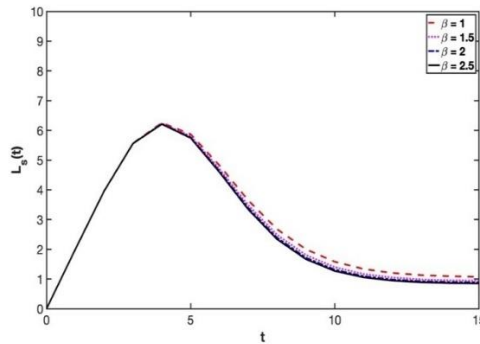
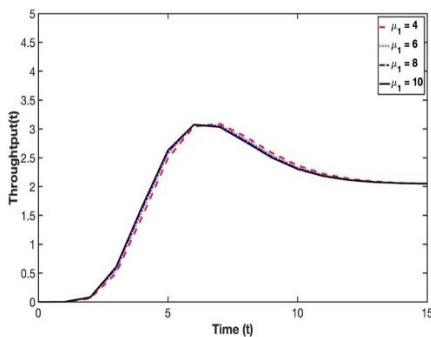
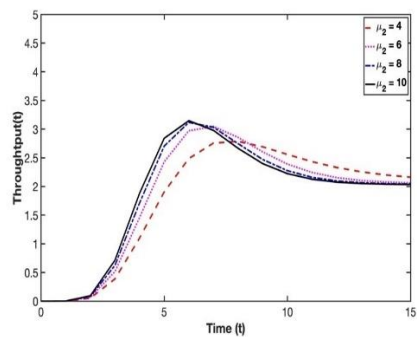


Figure 11: $L_s(t)$ vs μ_2

Figure 12: $L_s(t)$ vs ψ Figure 13: $L_s(t)$ vs α_2 Figure 14: $L_s(t)$ vs β

4.4 Effect on throughput

Throughput is measure of effectiveness of the queueing system. Effect of service rates on throughput is represented in figure 15-16, which shows that the number of effective services decreases for bigger values of service rates.

Figure 15: (t) vs μ_1 Figure 16: (t) vs μ_2

4.5 Effect on state probabilities

The sensitivity analysis of various state probabilities like probability of server being in idle state ($P_i(t)$), the probability of server being in busy state ($P_B(t)$) and the probability of server being in repair state ($P_R(t)$) for the different

values of arrival rate λ , repair rate β and time t are shown in table 1. The state probabilities are highly affected with different parameters on different time intervals as the data mentioned in table-1 for the obvious reasons.

Hence, a control on the parameters can truly be helpful in keeping a check on the performance of the system in various contexts. Like, N-policy is responsible for holding the customers before initiation of the services and in turn affect the cost of the system too. Similarly, repair policy helps in maintaining a balance between the repair services and arrival of the customers as system has finite space to hold upon the customers. Controlling or adjusting service rates for both the stages as per requirement is important as it influences the effective services of the system directly and its applicability to the concerned situation.

Table 1: Variation in State probabilities with different parameters

λ	Time (t)	$\beta = 2$			$\beta = 3$		
		$P_I(t)$	$P_B(t)$	$P_R(t)$	$P_I(t)$	$P_B(t)$	$P_R(t)$
2	5	0.2202	0.4601	0.0237	0.2201	0.4670	0.0165
	10	0.0078	0.8684	0.0243	0.0076	0.8753	0.0161
	15	0.0020	0.8813	0.0222	0.0019	0.8877	0.0148
3	5	0.0180	0.7258	0.0397	0.0180	0.7376	0.0270
	10	0.0052	0.8843	0.0344	0.0049	0.8938	0.0228
	15	0.0018	0.8859	0.0333	0.0018	0.8953	0.0222
4	5	0.0180	0.8146	0.0491	0.0008	0.8293	0.0333
	10	0.0044	0.8943	0.0456	0.0042	0.9069	0.0304
	15	0.0012	0.8941	0.0445	0.0010	0.9064	0.0297
5	5	0.0180	0.8502	0.0562	0.0180	0.8672	0.0382
	10	0.0032	0.9062	0.0563	0.0028	0.9222	0.0377
	15	0.0008	0.9061	0.0556	0.0008	0.9215	0.0371
6	5	0.0180	0.8693	0.0613	0.0180	0.8881	0.0419
	10	0.0018	0.9173	0.0642	0.0015	0.9363	0.0434
	15	0.0004	0.9185	0.0645	0.0001	0.9373	0.0435

5. Cost Analysis

Here, we frame the total cost function ($((t))$) for the system which uses various state probabilities and the cost analysis has been carried out. It is given as:

$$TC(t) = C_h L_s(t) + C_I P_I(t) + C_R P_R(t) + C_B P_B(t) + C_1 \mu_1 + C_2 \mu_2$$

where,

- C_h : holding cost per unit time for each customer present in the system;
 C_I : cost incurred due to idle system;
 C_R : repair cost;
 C_B : cost to be paid when the server is busy;
 C_1 : service cost for stage-1 services;
 C_2 : service cost for stage-2 services.

Following default values of the parameters have been taken for the calculation of total expected cost:

$$\lambda = 2, \mu_1 = 7, \mu_2 = 7, \alpha_1 = 0.5, \alpha_2 = 0.3, \beta = 2, \psi = 0.8, N = 8, q = 4, K = 30.$$

In addition, following set of values has been taken for cost analysis:

$$C_h = 30, C_I = 30, C_R = 20, C_B = 20, C_1 = 40, C_2 = 4.$$

The effect of various parameters with different values of time on total expected cost can be seen in figures 17-22.

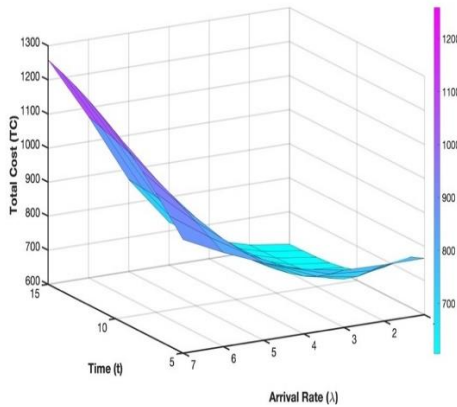


Figure 17: $TC(t)$ vs λ

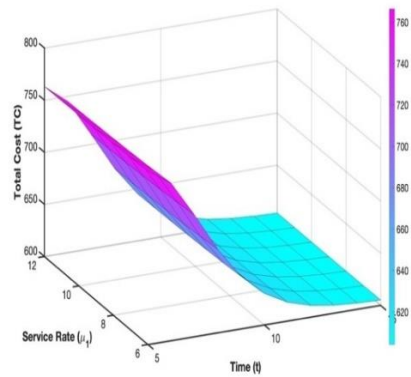


Figure 18: $TC(t)$ vs μ_1

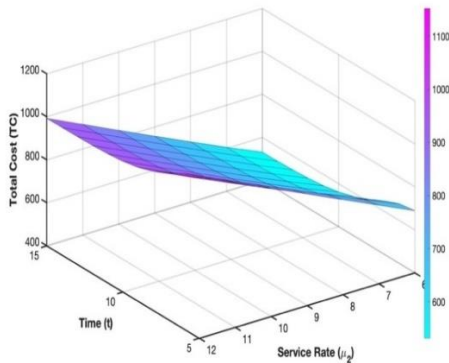


Figure 19: $TC(t)$ vs μ_2

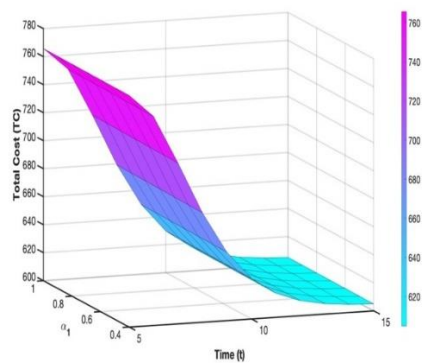


Figure 20: $TC(t)$ vs α_1

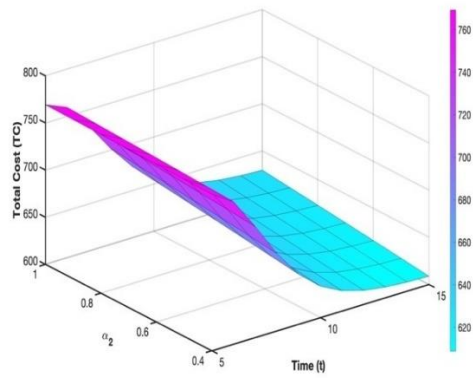


Figure 21: $TC(t)$ vs α_2

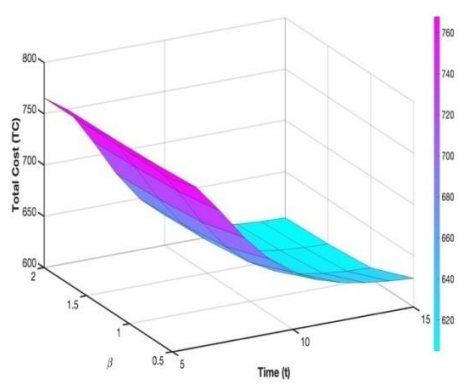


Figure 22: $TC(t)$ vs β

6. Conclusions

The present study investigates M/M/1 queueing model with threshold repair policy and N-policy working under the 2-stage compulsory services wherein the stage-2 service is offered only after completion of stage-1 service. The analysis of various vital performance measures gives a clear idea how the model behaves within certain limits and thus giving necessary information to concerned model designers and associated management about the how they can efficiently control the services and other parameters to reduce waiting times and queue length. The provision of two stage services also makes the system smoother by distributing the service load in two stages, thus saving the time and rush at last minute. This study is exciting and useful for the architects and management longing for design of such systems to accommodate a finite queue of customers waiting for services. This model is also significant for associated cost and time management team looking for ways to minimize the cost and also focusing on minimizing the waiting time of the customers by improving the services. The main idea behind the investigation was to analyse how the performance of a system is affected if instant repairing is reduced by threshold recovery, if single stage service is split into 2-stage and if N-policy is incorporated. It is clear from the results of throughput that indeed effective services are improved and queue length of the customers is also under control. Thus, managers are advised to implement such techniques for performance effective and cost effective outputs. The idea of customer feedback and reattempts can also be implemented in the model as future work that can keep a check on the servers and quality of service provided.

References

1. M. Yadin, P. Naor; Queueing systems with a removable service station, *Journal of the Operational Research Society*, 14 (4) (1963), 393-405.
2. D. P. Heyman; Optimal operating policies for M/G/1 queueing systems, *Operations Research*, 16 (2) (1968), 362-382.
3. J.-C. Ke, K.-H. Wang; A recursive method for the N policy G/M/1 queueing system with finite capacity, *European Journal of Operational Research*, 142 (3) (2002), 577-594.
4. D. Efrosinin, O. Semenova; An M/M/1 system with an unreliable device and a threshold recovery policy, *Journal of communications technology and electronics*, 55(12) (2010), 1526-1531.
5. M. Jain, G. Sharma, R. Sharma; Optimal control of (N, F) policy for unreliable server queue with multi-optional phase repair and start-up, *International Journal of Mathematics in Operational Research*, 4(2) (2012), 152-174.
6. M. Jain, A. Bhagat; Finite population retrial queueing model with threshold recovery, geometric arrivals and impatient customers, *Journal of Information and Operations Management*, 3(1) (2012), 162.
7. A. Kumar, M. Jain; Cost optimization of an unreliable server queue with two stage service process under hybrid vacation policy, *Mathematics and Computers in Simulation*, 204 (2023), 259-281.
8. G. Ayyappan, S. Karpagam; Analysis of a bulk queue with unreliable server, immediate feedback, N-policy, bernoulli schedule multiple vacation and stand-by server, *Ain Shams Engineering Journal*, 10(4) (2019), 873-880.
9. A. Ahuja, A. Jain, M. Jain; Transient analysis and ANFIS computing of unreliable single server queueing model with multiple stage service and functioning vacation, *Mathematics and Computers in Simulation*, 192 (2022), 464-490.
10. R. K. Meena, M. Jain, A. Assad, R. Sethi, D. Garg; Performance and cost comparative analysis for M/G/1 repairable machining system with n-policy vacation, *Mathematics and Computers in Simulation*, 200 (2022), 315-328.
11. S. R. Chakravarthy, R. Kulshrestha, et al.; A queueing model with server breakdowns, repairs, vacations, and backup server, *Operations Research Perspectives*, 7 (2020), 100131.