TUTA/IOE/PCU

*Journal of the Institute of Engineering, 2017, **13(1):** 60-68*

Printed in Nepal

Service Rate Optimization of Finite Population Queueing Model with State Dependent Arrival and Service Rates

Sushil Ghimire[1], Gyan Bahadur Thapa[2], Ram Prasad Ghimire[3]

[1,2] Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal

[3]Department of Mathematical Sciences, School of Science, Kathmandu University, Nepal

Corresponding author: *sushil198@gmail.com*

Received: *Oct 2, 2017* ***Revised:*** *Nov 27, 2017* ***Accepted:*** *Dec 5, 2017*

Abstract: Providing service immediately after the arrival is rarely been used in practice. But there are some situations for which servers are more than the arrivals and no one has to wait to get served. In this model, arrival rate is λ which follows a Poisson process and service time is exponentially distributed with rate μ. We have assumed the finite capacity queueing model and only N number of customers can get service. Customers more than N arrivals are rejected. We derive the explicit formulas for the average number of customers in the system by using recursive method to solve the system of steady state equations. Numerical results relevant to the performance indices have been presented so as to validate the results. The optimal rates of service have also been obtained by using routine optimization technique.

Key Words: Server, customer, optimization, performance, steady state.

1. Introduction

Standing in a queue in front of the bank, supermarket and many other places is our day to day experience. It is not an easy task to line up for the service. There are some special situations in which arrivals get the service instantly. It is a rare practice in queueing theory but also it is in use for some special situations in which waiting in a queue may result non-bearable loss. Everyone in a queue wants to be served first though it is not always possible. The common rule for service in a queue is first come first served (FCFS). There are some queueing disciplines in which last comer gets the service at first but in this paper, we have derived a model for which no arrivals have to wait for the service. No customers are expected to wait because servers are fixed more than the customers. This type of queueing model is known as infinite server queueing model. We are discussing here the Markovian queueing model which follows the Poisson arrival with rate λ and mean inter-arrival time is $1/\lambda$. Service time is distributed exponentially with rate μ so that mean service time is $1/\mu$. We are proposing this model only for the finite number of customers. Number of customers more than N will not be served. Another interest in this paper is to

optimize the service rate for the given values of arrivals to get the revenue. We have used the formula for the total expected cost and the total expected revenue for the calculation of total expected profit. We have not considered the constraints for the optimization model therefore routine method of calculus has been used to find the maximum revenue assuming the fixed value of arrival rate.

The paper is planned as follows: Section 2 describes the brief literature review. Section 3 includes notations used in the model together with the mathematical derivation and formula for mean number of customers in the system. Mathematics for the optimization of service rate is presented in Section 4. Section 5 describes the numerical results and interpretations. Finally, Section 6 concludes the paper.

2. Brief Literature Review

On arrival service for any customer in any queueing system is very difficult in practice. Queueing models of real life situations such as complex manufacturing system, transportation system, telecommunication system have to be tackled for the logical conclusion though they are expensive and hardly manageable. Increasing complexities of the queueing models, and vis- a-vis development of the techniques is due to the contributions of several researchers in the field. So, it is worthwhile to mention some of the works done on the line. Abidini et al. [1] studied a single-server multi-queue model for a vacation-type queueing system focusing mainly on queue length analysis. Ammar [2] derived expressions for the time dependent probabilities, mean and variance of the system size with some numerical illustrations to study the impatience customers and multiple vacations in a single server queueing system. Barache et al. [3] used $M/M/\infty$ queueing model to evaluate the stationary characteristics of the $GI/M/\infty$ queueing system and to observe the performance of the proposed model. Corral and Garcia [4] calculated maximum queue length during a fixed time interval using splitting methods and eigenvalue, eigenvector technique for an $M/M/c$ retrial queueing system. D'Auria [5] analysed $M/G/\infty$ queue to study the stochastic decomposition formula for the number of customers in the system with some examples in random environment. Ghimire et al. [6] verified formulas for mean queue length and mean waiting time using generating function technique for the batch arrival of customers. Ghimire et al. [7] calculated various performance measures for finite capacity time dependent multi-server queueing model and verified the results graphically using simulation. Gullu [8] considered $M/G/\infty$ queueing system for batch arrival in which same server serves the whole batch. Haviv and Oz [9] reviewed some existing observable queueing mechanisms where money transfers was taken into account concluding that the best ones are those in which customers have to make up their mind to join the queue without inspecting the queue length. Jiang et al. [10] dealt with a disaster $M/G/1$ queue in a multi-phase random environment in which the system stops working suddenly and resumes after exponential repair time. Kumar et al. [11] derived an optimization model of an $M/M/1/N$ feedback queue with retention of reneged customers. Roijers et al. [12] obtained all moments and covariance for congestion periods of a $M/M/\infty$ queue. Sah and Ghimire [13] studied transient Erlangian queueing system to calculate the different performance measures. Schweer and Wichelhaus [14] studied non-parametric $GI/G/\infty$ queueing system to estimate the service time distribution under partial information. Whitt [15] examined steady state infinite server queuing distribution where exponential service and sinusoidal arrival rate function is assumed. Wu et al. [16] examined the stability condition using

quasi-birth-and-death process for the optimization analysis of an unreliable multi-server queue with a controllable repair policy.

3. Mathematical Model

We derive the mathematical model of the problem by using the following notations:
$\frac{1}{\lambda}$ = mean inter-arrival time
$\frac{1}{\mu}$ = mean service time
L_s = mean number of customers in the system
W_s = mean waiting time in the system
C_s = cost of service per unit time
C_h = unit holding cost per unit time
C_L = cost associated with each lost unit
R = revenue earned by providing service to a customer
TEC = total expected cost per unit time of the system
TER = total expected revenue per unit time of the system
TEP = total expected profit per unit time of the system

A transient diagram is presented to establish the steady state balanced equations. Recursive method is used to solve the balanced equations to get mean number of customers in the system and the mean waiting time in the system.

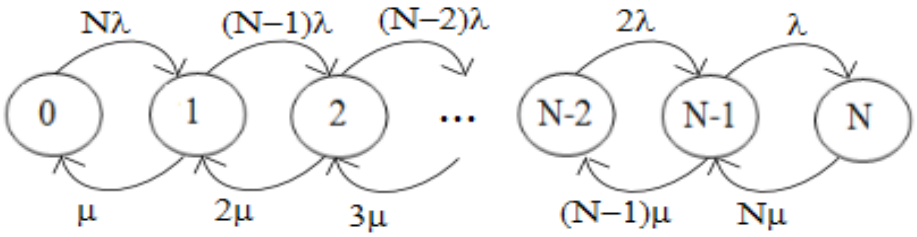

Fig 1: Transition diagram

Balanced equations are as follows:

$$N\lambda P_0 = \mu P_1 \quad (1)$$

$$(N-1)\lambda P_1 + \mu P_1 = N\lambda P_0 + 2\mu P_2 \quad (2)$$

$$(N-2)\lambda P_2 + 2\mu P_2 = (N-1)\lambda P_1 + 3\mu P_3 \quad (3)$$

continuing this way, we have

$$\lambda P_{N-1} + (N-1)\mu P_{N-1} = 2\lambda P_{N-2} + N\mu P_N \quad (4)$$

$$\lambda P_{N-1} = N\mu P_N \quad (5)$$

Solving these above equations

$$P_i = P_0 \left(\frac{\lambda}{\mu}\right)^i \binom{N}{i} \tag{6}$$

Now, the probability normalizing conditions is $\sum_{i=0}^{N} P_i = 1$

Substituting the value of P_i from equation (6)

$$\sum_{i=0}^{N} P_0 \left(\frac{\lambda}{\mu}\right)^i \binom{N}{i} = 1$$

$$\Longrightarrow \quad P_0 = \frac{1}{\left(1+\frac{\lambda}{\mu}\right)^N}$$

With this value of P_0 expression for P_i is

$$P_i = \begin{cases} \dfrac{\left(\frac{\lambda}{\mu}\right)^i \binom{N}{i}}{\left(1+\frac{\lambda}{\mu}\right)^N} & 0 \le i \le N \\ 0 & \text{otherwise} \end{cases}$$

Let L_s denote the average number of customers in the system then,

$$L_s = \sum_{i=0}^{N} iP_i$$

After suitable simplification, we have

$$L_s = \frac{N.\frac{\lambda}{\mu}}{\left(1+\frac{\lambda}{\mu}\right)}$$

In this model, we propose that no customer has to wait for the service. In this scenario, mean number of customers in the queue and mean waiting time in the queue are both zero. Mean waiting time in the system is almost the same as the average service time.

Therefore, $W_s = \frac{1}{\mu}$.

4. Optimization Model

In this Section, we calculate the maximum service capacity per unit time for the total expected profit in the system. We use the MATLAB simulation to get the values of the maximum service rate per unit time for the different values of arrival rate per unit time.

The total expected cost (TEC) per unit time of the system is given by

TEC = $\mu C_s + C_h L_s + C_L \lambda P_N$

$$= \mu C_s + C_h \frac{N.\frac{\lambda}{\mu}}{\left(1+\frac{\lambda}{\mu}\right)} + C_L \lambda \frac{\left(\frac{\lambda}{\mu}\right)^N}{\left(1+\frac{\lambda}{\mu}\right)^N}$$

Again, the total expected revenue (TER) per unit time of the system is given by

TER = $R\mu(1 - P_0)$

$$= R\mu\left(1 - \frac{1}{\left(1+\frac{\lambda}{\mu}\right)^N}\right)$$

Now, the total expected profit (TEP) is

TEP = TER –TEC

$$= R\mu\left(1 - \frac{1}{\left(1+\frac{\lambda}{\mu}\right)^N}\right) - \mu C_s - C_h \frac{N.\frac{\lambda}{\mu}}{\left(1+\frac{\lambda}{\mu}\right)} - C_L \lambda \frac{\left(\frac{\lambda}{\mu}\right)^N}{\left(1+\frac{\lambda}{\mu}\right)^N}$$

$$= R\mu - R\frac{\mu^{N+1}}{(\mu+\lambda)^N} - \mu C_s - C_h \frac{N.\lambda}{(\mu+\lambda)} - C_L \frac{\lambda^{N+1}}{(\mu+\lambda)^N}$$

From the point of view of optimization of the problem, the objective function is

$$\text{Maximize: TEP} = R\mu - R\frac{\mu^{N+1}}{(\mu+\lambda)^N} - \mu C_s - C_h \frac{N.\lambda}{(\mu+\lambda)} - C_L \frac{\lambda^{N+1}}{(\mu+\lambda)^N}$$

Differentiating TEP partially with respect to μ, we have

$$\frac{\partial(\text{TEP})}{\partial \mu} = R - R\frac{(N+1).(\mu+\lambda)^N.\mu^N - N.\mu^{N+1}.(\mu+\lambda)^{N-1}}{(\mu+\lambda)^{2N}} - C_s + C_h \frac{N.\lambda}{(\mu+\lambda)^2} + C_L \frac{N.\lambda^{N+1}}{(\mu+\lambda)^{N+1}}$$

$$= R - R\mu^N\left(\frac{(N+1).-N\mu.(\mu+\lambda)^{-1}}{(\mu+\lambda)^N}\right) - C_s + C_h \frac{N.\lambda}{(\mu+\lambda)^2} + C_L \frac{N.\lambda^{N+1}}{(\mu+\lambda)^{N+1}}$$

$$= R - \frac{R(N+1)\mu^N}{(\mu+\lambda)^N} + \frac{RN\mu^{N+1}}{(\mu+\lambda)^{N+1}} - C_s + C_h \frac{N.\lambda}{(\mu+\lambda)^2} + C_L \frac{N.\lambda^{N+1}}{(\mu+\lambda)^{N+1}}$$

Again differentiating partially with respect to μ

$$\frac{\partial^2(\text{TEP})}{\partial \mu^2} = -R(N+1)\left(\frac{(\mu+\lambda)^N.N.\mu^{N-1} - \mu^N.N.(\mu+\lambda)^{N-1}}{(\mu+\lambda)^{2N}}\right) +$$
$$RN\left(\frac{(\mu+\lambda)^{N+1}.(N+1).\mu^N - \mu^{N+1}.(N+1).(\mu+\lambda)^N}{(\mu+\lambda)^{2(N+1)}}\right) - \frac{2C_h N.\lambda}{(\mu+\lambda)^3} - \frac{C_L N.(N+1)\lambda^{N+1}}{(\mu+\lambda)^{N+2}}$$

$$= -R(N+1)\left(\frac{N.\mu^{N-1} - \mu^N.N.(\mu+\lambda)^{-1}}{(\mu+\lambda)^N}\right) + RN\left(\frac{(N+1).\mu^N - \mu^{N+1}.(N+1).(\mu+\lambda)^{-1}}{(\mu+\lambda)^{(N+1)}}\right) - \frac{2C_h N.\lambda}{(\mu+\lambda)^3} -$$
$$\frac{C_L N.(N+1)\lambda^{N+1}}{(\mu+\lambda)^{N+2}}$$

We have calculated the maximum number of service capacity for the different revenue and different arrival rates. If the arrival rate is $\lambda = 4$ and the revenue is to be 100, the maximum service capacity is approximately 42. Likewise, for the revenue to be 20, the maximum service capacity is approximately 15. This result indicates that more the revenue is more the profit will be, which is realistic in nature.

Table 1: Optimal service rates for C_s=4, C_h=3, C_L=8, N=3

R	$\lambda = 3$	$\lambda = 4$	$\lambda = 5$
	μ^*	μ^*	μ^*
100	31.53	42.07	52.62
70	25.48	34.02	42.56
60	23.18	30.95	38.73
50	20.67	27.61	34.56
30	14.72	19.70	24.68
20	10.94	14.69	18.43

Table 1 shows the values of maximum service rates for the given values of arrivals to get the given revenue. We have taken three different values of λ so as to get the revenue and the maximum service rate.

5. Numerical Results and Interpretations

MATLAB simulation has been used to verify the model. Fig. 2 is the graph for mean number of customers against arrival rate which indicates that number of customers in the system increases for the bigger arrival rate. For the smallest arrival rate, the graph is close to the x-axis and for the higher service rate the number of requests decreases in the system.

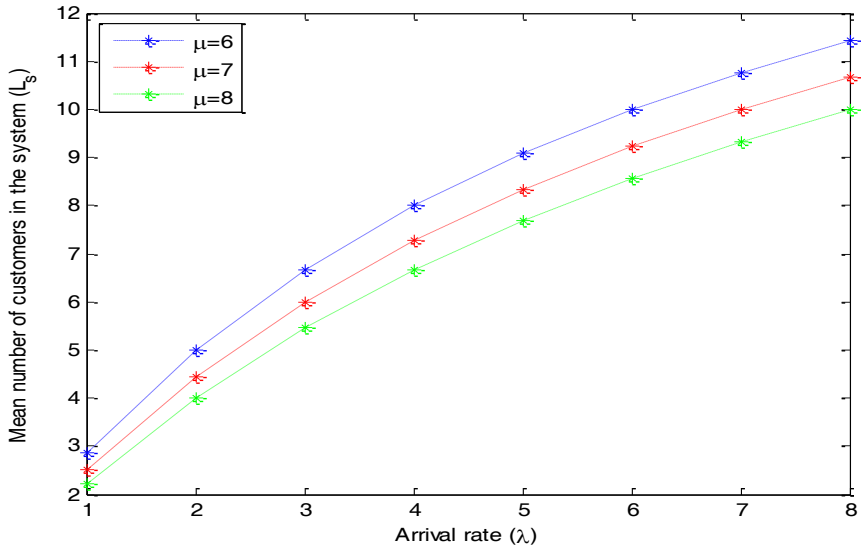

Fig. 2: Arrival rate *vs.* mean number of customers in the system

On the other hand, Fig. 3 is the graph for mean number of customers vs. service rate. We can see that, for more service rate there are less number of customers in the system. For less arrival rate, less number of customers and for more arrival rate more number of customers in the system has been observed indicating that the model we established is appropriate.

The graph at the bottom is for the least arrival rate 6 which indicates the less number of customers in the system whereas the numbers of customers are gradually increasing in the other two graphs for the arrival rates 7 and 8 respectively.

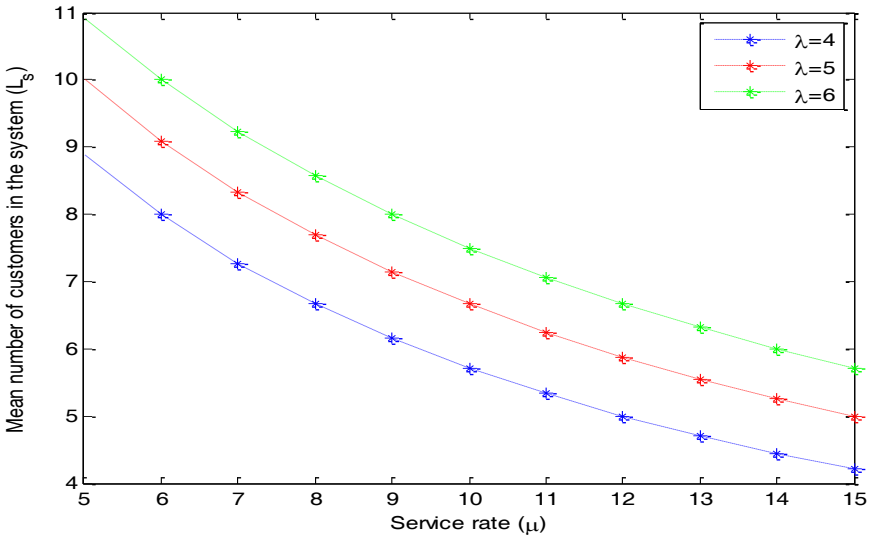

Fig. 3: Service rate vs. mean number of customers in the system

These two graphs obtained by using the MATLAB software are important to see the real applicability of the model in everyday life. Different values of arrival rate and service rate has been taken just to see the increased or decreased pattern of request in the system.

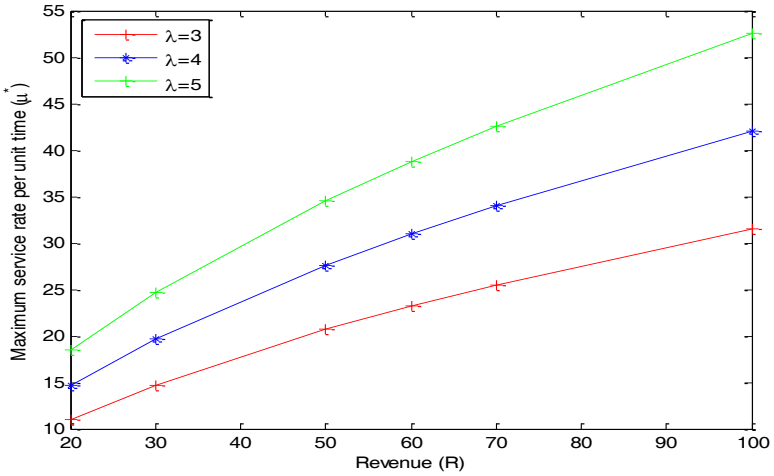

Fig 4: Revenue vs. maximum service rate per unit time

Fig. 4 is plotted revenue against maximum service rate per unit time. This shows that more revenue can be obtained if there is more service rate per unit time. It is seen from the different curves that for the more arrival rate service rate should be increased which finally results the more revenue.

6. Conclusion

In this model customers do not need to wait for the service. Though this type of queueing model is rarely seen in practice, it is an interesting part in the study of queueing theory. All the arrivals get the service at the time they come for service. We have calculated the average number of customers in the system and plotted the graph for it against arrival rate and service rate. It is the finite population queueing model so the arriving customers exceeding N cannot get service. If the system capacity is considered unlimited the study becomes more interesting and challenging. Moreover, including customers' behaviour like balking, reneging or jockeying makes the model more realistic.

Providing service for the high class customers is one of the examples of this model. For an example, some of the telephone companies set number of towers so that no calls fail. During rescue operations, one server is supposed to rescue one individual. In some special occasions, restaurant prepares the menu at the table before the customers' arrival. Moreover, some internet providers manage a very high speed for some special persons to avoid complain from them. The model under study can be experienced in processor sharing queueing system also where all the customers get the service at a time like wild animals share the same river.

Acknowledgement: The first author is thankful to Erasmus Mundus SmartLink project for financial support as PhD exchange student to visit in Burgas Free University, Bulgaria, Sep 2016 – Sep 2017. The second author is thankful to Erasmus Mundus LEADER Project for funding as Post Doc Research Fellow in Department of Mathematics, University of Evora, Portugal, Nov 2016 - Aug 2017.

References

[1] Abidini MA, Boxma O and Resing J (2016), Analysis and optimization of vacation and polling models with retrials. *Performance Evaluation*, **98:** 52–69.

[2] Ammar SI (2015), Transient analysis of an $M/M/1$ queue with impatient behaviour and multiple vacations. *Applied Mathematics and Computation*, **260:** 97–105.

[3] Bareche A, Cherfaoui M and Aissani D (2016), Approximate analysis of an $GI\ /M/\infty$ queue using the strong stability method. *IFAC-PapersOnLine*, 49-12, 863–868.

[4] Corral AG and Garcia ML (2014), Maximum queue lengths during a fixed time interval in the $M/M/c$ retrial queue. *Applied Mathematics and Computation*, **235:** 124–136.

[5] D'Auria B (2007), Stochastic decomposition of the $M/G/\infty$ queue in a random environment. *Operations Research Letters*, **35:** 805 – 812.

[6] Ghimire S, Ghimire RP and Thapa GB (2014), Mathematical models of $M^b/M/1$ bulk arrival queueing system. *Journal of the Institute of Engineering*, **10(1):** 184-191.

[7] Ghimire S, Ghimire, RP and Thapa GB (2015), Performance evaluation of unreliable $M(t)/M(t)/n/n$ queueing system. *British Journal of Applied Science & Technology*, **7(4):** 412-422.

[8] Gullu R (2004), Analysis of an $M/G/\infty$ queue with batch arrivals and batch-dedicated servers. *Operations Research Letters*, **32:** 431–438.

[9] Haviv M and Oz B (2016), Regulating an observable $M/M/1$ queue. *Operations Research Letters*, **44:** 196–198.

[10] Jiang T, Liu L and Li J (2015), Analysis of the $M/G/1$ queue in multi-phase random environment with disasters. *J. Math. Anal. Appl.*, **430:** 857–873.

[11] Kumar R, Jain NK and Som BK (2014), Optimization of an M/M/1/N feedback queue with retention of reneged customers. *Operations Research and Decisions*, **3:** 45-58.

[12] Roijers F, Mandjes M and Berg HVD (2007), Analysis of congestion periods of an $M/M/\infty$ queue. *Performance Evaluation*, **64:** 737–754.

[13] Sah SS and Ghimire RP (2015), Transient analysis of queueing model. Journal of the Institute of Engineering. **11(1):** 165-171.

[14] Schweer S and Wichelhaus C (2015), Nonparametric estimation of the service time distribution in the discrete-time $GI/G/\infty$ queue with partial information. *Stochastic Processes and their Applications*, **125:** 233–253.

[15] Whitt W (2014), The steady-state distribution of the $M_t/M/\infty$ queue with a sinusoidal arrival rate function. *Operations Research Letters*, **42:** 311–318.

[16] Wu CH, Lee WC and Ke JC (2014), Optimization analysis of an unreliable multi-server queue with a controllable repair policy. *Computers & Operations Research*, **49:** 83**–**96.