APPLICATION OF QUEUING THEORY FOR THE IMPROVEMENT OF BANK SERVICE

[1]TOSHIBA SHEIKH, [2]SANJAY KUMAR SINGH, [3]ANIL KUMAR KASHYAP

[1,2,3]Chhatrapati Shivaji Institute of Technology, Durg (C.G.), India
Email: tsheikh1990@gmail.com[1], sanjaysingh@csitdurg.in[2], anilkashyap@csitdurg.in

Abstract— Lines of waiting customers are always very long in most of the banks. In this paper, the M/M/Z/∞:FCFS model is converted into M/M/1/∞:FCFS to know which one is more efficient, a line or more lines. To do this, first we establish the optimization model of queuing and calculate the optimal model of queuing. Second, calculate the optimal number of service stations to improve operational efficiency. Third, we calculate the optimal service rate and the service efficiency by the operating costs. Based on these aspects, the result of analysis was effective and practical.

Keywords— Queuing Model, Optimal service stations, Optimal service rate, M/M/z/∞ Model, First Come First Serve.

I. INTRODUCTION

One of the serious classes of queuing systems that we all encounter in our daily lives is commercial service systems, where customers receive service from commercial organizations. Many organizations involve person-to-person service at a fixed location, such as a barber shop, cafeteria, petrol pump and bank. Many commercial banks have done great effort to increase the service efficiency and customer satisfaction but the most of them are facing a serious problem of waiting line of customers. In bank, the waiting line of customers appears due to low efficiency of the queuing system, it reflects the lacking of the business philosophy of customer centric, low service rate of the system. The waiting queues of the customer develop because the service to a customer may not be delivered immediately as the customer reaches the service facility [8]. Lack of satisfactory service facility would cause the waiting line of customers to be formed. The only technique is that the service demand can be met with ease is to increase the service capacity and increasing the efficiency of the existing capacity to a higher level. In the following, to solve the problem of the long waiting lines of the customer is studied by means of the queuing theory, the determination to reduce the time of customers waiting is obtained to achieve the goal of people oriented and the greatest effectiveness of the banks.

II. QUEUING THEORY AND MODEL FORMULATION

A. Queuing Theory:

Queuing theory is basically a mathematical approach applied to the analysis of waiting lines. It uses models to represent the various types of queuing systems. Formula for each model indicates how the related queuing system should perform, under a variety of conditions. The queuing model are very powerful tool for determining that how to manage a queuing system in the most effective manner [9]. The queuing theory is also known as the random system theory, which studies the content of: the behavior problems, the optimization problem and the statistical inference of queuing system [4].

B. Terminology and Notations:

The following terminology and notations are used in the model formulation and calculations:

P_n= probability of exactly n customers in the system.
N= number of customers in the system.
L_s= expected number of customers in the system
L_q= expected number of customers in the queue.
W_s= waiting time of customers in the system
W_q= waiting time of customers in the queue.
λ_n= The mean arrival rate (expected number of arrivals per unit time) of new customers are in systems.
μ_n= The mean service rate for overall systems (expected number of customers completing service per unit time) when n customers are in systems [4][6].

The mean arrival rate is constant for all n, this is denoted by λ and the mean service rate per busy server is constant for all n≥1, is denoted by μ. And when n≥z that is all z servers are busy, μ=sμ. Under this condition, the expected inter-arrival time is 1/λ and the expected service time is 1/μ. The utilization factor for the service facility is ρ= λ/μz , i.e., the expected fraction of time as the individual servers are busy, because λ/μz represents the fraction of the system's service capacity (zμ) that is being utilized on the average by arriving customers λ [8].

C. M/M/1/∞ Model:

This is the simplest queuing system to analyze. The system consists of only one server. The arrivals follow Poisson distribution with a mean arrival rate of λ and the service time has exponential distribution with the average service rate of μ. $p_n = p(N=n), (n=0,1,2,\ldots)$ is the probability distribution of the queue length.

Utilization factor i.e. the fraction of time servers are busy:

$$\rho = \frac{\lambda}{\mu}$$

Expected number of customers in the system

$$L_s = \frac{\rho}{1-\rho}$$

Expected number of the customers waiting on the queue:

$$L_q = \frac{\rho^2}{1-\rho}$$

Expected waiting time of customers in the queue:

$$W_q = \frac{\rho}{\mu - \lambda}$$

Expected waiting time of customers in the queue:

$$W_s = \frac{1}{\mu - \lambda}$$

D. M/M/Z/∞ Model:

This model treats the condition in which there are several service stations in parallel and each customer in the waiting queue can be served by more than one station channel. Consider an M/M/z queue with arrival rate λ, service rate μ and z servers. The traffic intensity is defined usual by the ratio $\rho_z = \frac{\rho}{z} = \frac{\lambda}{z\mu}$.

The steady distribution of queuing system [4] is studied as following:

$p_n = p(N=n), (n=0,1,2,\ldots)$ is the probability distribution of the queue length *N*, as the system is in steady state, when the number of system servers is Z, then we have $\lambda_n = \lambda$, n=0,1,2,…

If there are n customers in the queuing system at any point in time, then the following two cases may arise:

1. If n<Z, (number of customers in the system is less than the number of servers), then there will be no queue. However, (Z–n) number of servers will not be busy. The combined service rate will then be $\mu_n = n\mu$; n<Z.
2. If n≥Z, (number of customers in the system is more than or equal to the number of servers) then all servers will be busy and the maximum number of customers in the queue will be (n – Z). The combined service rate will be $\mu_n = z\mu$; n≥Z.

From the model the probability of having n customers in the system is given by:

$$\rho_z = \lambda / z\mu$$

$$p_0 = \left[\sum_{n=0}^{z-1} \frac{1}{n!} \left(\frac{\lambda}{\mu} \right)^n + \frac{1}{z!} \left(\frac{\lambda}{\mu} \right)^z \frac{z\mu}{z\mu - \lambda} \right]^{-1}$$

$$p_n = \begin{cases} (\rho^n / n!)\, p_0 & n \le z \\ \rho^n / (z! z^{n-z})\, p_0 & n > z \end{cases}$$

When n≥Z, it is that the number of customers in the system is not smaller than the number of servers, the next customers must wait, that is,

$$C(z,\rho) = \sum_{n=z}^{\infty} P_n = \frac{\rho^z}{z!(1-\rho_z)} P_0$$

Where $\rho_z = \rho / z = \lambda / (z\mu)$

Expected number of the customers waiting on the queue:

$$L_q = \left[\frac{1}{(z-1)!} \left(\frac{\lambda}{\mu} \right)^z \frac{\mu\lambda}{(z\mu - \lambda)^2} \right] p_o$$

Expected number of customers in the system:

$$L_s = L_q + \frac{\lambda}{\mu}$$

Expected waiting time of customers in the queue:

$$W_q = \frac{L_q}{\lambda}$$

Expected time a customer spends in the system:

$$W_s = \frac{L_s}{\lambda}$$

III. OPTIMIZATION IN BANK QUEUE

By means of the queuing theory, the bank queuing problem is studied as the following aspects:

A. *One Line or More:*

In reality we have waiting lines in the bank, there are several service stations. Each service station has a queue or a waiting line. If each service station has a queue according to their schedule, the arrival customers join in each queue as the probability 1/2, known as the two scheduled queue. For example, when there are two lines, the system can be considered as two isolated M/M/1 systems, and the arrival rate of each service station λ=λ/2. If there is a line, the system will be M/M/2, L, L_q, W and W_q are calculated respectively and compared to know which one is more efficient, we will analysis it from a technical point as following:

When there is a line, z=2, λ=50, μ=40, ρ=5/4

$$p_0 = \left[\sum_{n=0}^{z-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{z!}\left(\frac{\lambda}{\mu}\right)^z \frac{z\mu}{z\mu-\lambda} \right]^{-1} = 0.230$$

$$L_q = \left[\frac{1}{(z-1)!}\left(\frac{\lambda}{\mu}\right)^z \frac{\mu\lambda}{(z\mu-\lambda)^2} \right] p_o = 0.801$$

$$L_s = L_q + \frac{\lambda}{\mu} = 2.051$$

$$W_s = \frac{L_s}{\lambda} = 0.041$$

$$W_q = \frac{L_q}{\lambda} = 0.016$$

When there are two lines, λ=λ/2=25, μ=40, ρ=5/8

$$L_s = \frac{\rho}{1-\rho} = 1.667$$

$$L_q = \frac{\rho^2}{1-\rho} = 1.041$$

$$W_s = \frac{1}{\mu-\lambda} = 0.067$$

$$W_q = \frac{\rho}{\mu-\lambda} = 0.041$$

Similarly, the calculation is same for, when the lines are three, four and five.

When there are n lines: it is means that there are n service stations, each service station has a queue based on their schedule, each arrival customer joins in each queue at the probability 1/n. It is called as scheduled n queues. The mean arrival rate is λ/n, the mean service rate is μ.

Expected number of customers in the system:

$$L_s = \frac{\lambda/n}{\mu-\lambda/n} = \frac{\lambda}{n\mu-\lambda}$$

Expected number of the customers waiting on the queue:

$$L_q = \frac{\lambda L}{n\mu} = \frac{\lambda^2}{n\mu(n\mu-\lambda)}$$

Average time a customer spends in the system:

$$W_s = \frac{1}{\mu-\lambda/n} = \frac{n}{n\lambda-\lambda}$$

Expected waiting time of customers in the queue:

$$W_q = \frac{nL_q}{\lambda} = \frac{\lambda}{\mu(n\mu-\lambda)}$$

From the Table I, we can see that, in the case of two lines, the waiting time in the system is 0.067 and at a line, it is 0.041. The staying time is decreasing clearly, and the length of line is also less. This shows that in banking services, in terms if "first come, first serve" the principle of fairness or technically, a line is better than more lines, so bank managers should have the attention on this problem.

Table I. Main characteristics in the queuing systems

Num	λ	μ	L_s	L_q	W_s	W_q
1	50	40	2.051	0.801	0.041	0.016
2	25	40	1.667	1.041	0.067	0.041
3	16	40	0.714	0.297	0.042	0.017
4	12	40	0.454	0.142	0.036	0.011
5	10	40	0.333	0.083	0.033	0.008

B. *Optimal Service Station:*

In order to guarantee the quality of service, we set up the number of service stations. If we need customers need to line up no more than 10% how many service stations should be setup.

When z=2, λ=50, μ=40, ρ=5/4

$$p_0 = \left[\sum_{n=0}^{z-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{z!}\left(\frac{\lambda}{\mu}\right)^z \frac{z\mu}{z\mu-\lambda} \right]^{-1} = 0.230$$

$$C(z,\rho) = \sum_{n=z}^{\infty} P_n = \frac{\rho^z}{z!(1-\rho_z)} P_0$$

=0.288

Similarly for the stations 3, 4 and 5 the calculations are same as for the 2 stations. We can see from the

Table II, when the number of service stations is 4; the probability of queuing is 3.86% less than 10%. Based on this, bank managers could set the propriety service station to improve service.

Table II. Number of service stations verses the probability of queuing

Number of service stations	2	3	4	5
The probability of queuing	28.84%	15.54%	4.22 %	0.97%

C. Optimal Service Rate:

Here we only studied the condition of one service station that is; consider the model M/M/1/∞. To determine the particular level of service, which minimizes the total cost of providing service and waiting for that service.

Let C_w = expected waiting cost/unit/unit time.
L_s = expected (average) number of units in the system
C_s = cost of servicing one unit.

Expected waiting cost per unit time,

$$C_w.L_s = C_w.\frac{\lambda}{\mu-\lambda}$$

Expected service cost per unit time is, $C_s.\mu$

Total cost, $C = C_w \frac{\lambda}{\mu-\lambda} + \mu.C_s$

This will be minimum if $\frac{dC}{d\mu} = 0$

We have, $C_s - C_w \frac{\lambda}{(\mu-\lambda)^2} = 0$

The optimal service rate is: $\mu^* = \lambda + \sqrt{\frac{C_w}{C_s}\lambda}$

With which we can find out the optimal services, to improve the efficiency of our services [8].

CONCLUSION

The efficiency of commercial banks is improved by the following three measures: the queuing number, the service stations number and the optimal service rate are investigated by means of queuing theory. By the example, the results are effective and practical. The time of customer queuing is reduced. The customer satisfaction is increased. It was proved that this optimal model of the queuing is feasible.

REFERENCES

[1] Qun Zhang & Zhonghui Dong Zhian, "Dynamic Optimization of Commercial Bank Major Channels," Fourth International Conference on Business Intelligence and Financial Engineering, pp. 627-630, 2011.

[2] Yu-Bo WANG, Cheng QIAN and Jin-De CAO "Optimized M/M/c Model and Simulation for Bank Queuing System", IEEE, 2010.

[3] Olaniyi, T. A., "An Appraisal of Cost of Queuing in Nigerian Banking Sector: A Case Study of First Bank of Nigeria Plc, Ilorin," Journal of Business and social sciences. Vol. 9, No. 1&2, pp. 139-145, 2004.

[4] H. M. Xiao, Z. C. Zang, Y. X. Zhang, "Operational Research," Chengdu, China, Electronic Science and Technology University Press, 2003.

[5] Operational Research materials group, operational Research, Beijing, China, Tsinghua University Press 2005.

[6] Y. H. Guo, X. P. Zhinh, The queuing model of dynamic vehicle routing problem, Journal of Management Science, Vol. 9, No. 1, pp. 33-37, 2006.

[7] H. Zeng, Queue based on the cost of distribution systems, Journal of Wuhan University of Science and Technology, Vol. 30, No. 3, pp. 330-332, 2007.

[8] F. S. Hillier and G. J. Lieberman, "Introduction to Operations Research," Tata Mc-Graw Hill, 8th Edition, 2007.

★★★