Journal of Modern Science and Technology
Vol. 4. No. 1. September 2016 Issue. Pp. 144 – 153

Cost Effective Decisions on Service Rate at a Customer Care Centre with Multi-Server Queuing Model

Khairun Nahar*, Md. Zahidul Islam** and Md. Minarul Islam***

In this study the queuing characteristics of Teletalk Customer Care Centre, Rajshahi is observed and analyzed with a multi-server queuing model. The waiting and service costs are also analyzed to determine the optimal service rate and the service rate is utilized to measure the number of servers (Customer Care Officer). TORA optimization software is utilized to identify and analyze the results. Other important performance characteristics (average waiting time of the customer in the system, average no. of people in the system etc.) of the queuing system are also analyzed for different number of server. The existing system is working with 5 servers but it costs high significantly. In this study the optimal number of server for the system is identified with optimum total costs (waiting and service costs).

Keywords: Multi-server, Queuing, Cost

1. Introduction

Queues (waiting lines) are a part of everyday life. It is a natural phenomenon (Hiller & Liberman, 2001). We all wait in queues to buy a Bus ticket, make a bank deposit, pay for groceries, mail a package, obtain food in a cafeteria etc. We have become accustomed to considerable amounts of waiting, but still get annoyed by unusually long waits. However, having to wait is not just a petty personal annoyance. The amount of time that a nation's populace wastes by waiting in queues is a major factor in both the quality of life there and the efficiency of the nation's economy. For example, before its dissolution, the U.S.S.R. is notorious for the tremendously long queues that its citizens frequently had to endure just to purchase basic necessities. Even in the United States today, it has been estimated that Americans spend 37,000,000,000 hours per year waiting in queues. If this time could be spent productively instead, it would amount to nearly 20 million person-years of useful work each year (Hiller & Liberman, 2001)!

So, we can say that a very good queue management system is needed to compensate this huge amount of loss. Queues usually grow when the supply of a service is less than the demand (Panneersalvam, 2009). The length of a queue largely depends upon the number of customer arriving at the service facility and the service rate. Wait time depends on the number of people in the system, number of service line (servers) and amount of time needed to serve a customer (Cooper, 1981). In a customer care Centre people come to take service and they don't want to wait in a queue. Queuing system is designed for a particular number of servers; the number of server is determined as a way to minimize the total cost to run a system

*Khairun Nahar, Corresponding Author, Assistant Professor, Department of Industrial & Production Engineering, Rajshahi University of Engineering & Technology, Rajshahi, Bangladesh,
Email: shapla05.ipe@gmail.com
**Md. Zahidul Islam, Department of Industrial and Production Engineering (IPE), Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh, Email: zahid.ruet11@gmail.com
***Md. Minarul Islam, Department of Industrial and Production Engineering (IPE), Rajshahi University of Engineering and Technology, Rajshahi 6204, Bangladesh.

(Gorunescu, Mc Clean & Millard, 2002). There are basically two types of costs associated with a queuing system they are Service Costs and Waiting Costs associated with the customer (Adan & Resung, 2002). Keep waiting customers in a queue can be very dangerous, they might not willing to come back again to take the service or also they can switch the operator (Kembe, Onah, & Lorkegh, 2012). A queuing system can be set up with excessive amount of service facility which is not optimal for this system, this will led to unnecessary cost which is not required for running the system. The goal of queuing theory is therefore to minimize the total cost to the system. The two costs mentioned earlier are very important to make a decision about an existing queuing system. There is an inverse relation between the two costs as the Service Costs increase with an increase in number of servers while the Waiting Costs decrease with an increase in number of servers (Hiller & Liberman, 2001). Waiting costs may include the loss of future business due to the dissatisfaction of the customer, the loss of the customer due to wait in the queue (opportunity cost); he could use this time in other effective work. The quality of service can be decreased by too much waiting in a queuing system. On the other hand Service Cost is the cost of providing service .These includes salaries paid to employees or servers while they wait for giving service to customers. Cost of waiting space, facilities, equipment, and supplies. Managers can take decisions about a queuing system by minimizing the total costs to run a system. Queues or waiting lines or queuing theory, is first analyzed by A.K. Erlang a Danish Engineer in 1913 in the context of telephone facilities (Hiller & Liberman, 2001). He is experimenting with the fluctuating demand for telephone facilities and its effect on automatic dialing equipment at the Copenhagen telephone System. Since World War II this theory has been applied to many business and human service fields. Literatures on queuing indicate that waiting in line or queue causes inconvenience to economic costs to individuals and organizations such as healthcare, airline companies, banks, manufacturing firms etc. But there is no research found that ascertain the optimum number of servers considering different costs for example waiting, service costs etc. as mentioned above for telecom company (in this case the teletalk customer care centre at rajshahi, Bangladesh). In this paper corresponding service cost and waiting cost is calculated according to the system of this particular case. Service cost is calculated as the hourly marginal cost of a server and the waiting cost of a customer by using the per capita annual income of the people of Bangladesh. This study provides optimum number of server of customer care center service facility by minimizing total service and waiting cost.

This paper arranged as follows- Section 2 presents the literature review to identify the research scope of implementing queuing theory in telecom sector. Section 3 explains the methodology, queuing model and cost analysis. In this section, the Multi-server queuing model is explained in detail which is utilized to solve the queuing problem of a telecom company of Bangladesh and it is elaborated as a case study. Section 4 presents data analysis and findings, and Section 5 presents the conclusions of the study.

2. Literature Review

Literature review works as a guideline to select a problem area and justifies the approaches taken to solve the problems. We have studied many books and journal papers to enrich our knowledge about queuing theory. Queuing theory is first developed by A. k Erlang a Danish engineer at 1913. The theory is then developed by many scientists. There are many way that queuing theory can be used, Kembe (2012), McClean (2002), Troy (2011) used queuing theory to develop the service system of a

hospital. They also developed model to determine the service level of the hospitals. Service level means the number of doctor and number of beds required to run the queue system of a hospital. Queuing theory is also used in different other fields such as in banking sector, Rahman (2013) and Olaniyi (2004) also used the knowledge of queuing theory to develop and describe the characteristics of the queue system in banking sector. Different types of queue model are used by different authors to fit different criteria at different circumstances. They also developed different system by different model. Rahman (2013) and Kembe (2012) used Multi-Server queuing model in their research. McClean (2002) used another criterion for bed-occupancy management and planning of hospital. Different authors describe queuing theory in their book differently. Liberman (2001), Taha (2007), Panneersalvam (2009) Cooper (1981) and Resung (2002) described queuing theory in their book differently. They showed enormous amount of practical cases in their books. They also showed different practical applications of queuing theory and different theories and models. To determine the optimal service level for a system introducing the cost into the model is very important. The main objective is to minimize the total cost of a system. There are basically two types of cost in a system they are Service Cost and Waiting cost by the customers. Different researcher had shown different techniques to calculate these costs. They all basically calculated the service cost as an hourly cost or as a cost in a certain time period. They all considered marginal service cost in a time period as the service cost. Different people calculated waiting cost in different way. Kembe (2012) used waiting cost but did not show the way to calculate it. Rahman (2013) calculated waiting cost by using probability. He divided the population in different socio professional category and had done a survey on them. Then he calculated the mean hourly income of the population by using those data ranges. Liberman (2001) showed in his book the relationship between the service cost and waiting cost in a queuing system to determine the optimal service rate. He also described the way to develop the function of waiting cost. We have used in our study Multi-server queuing model (Kembe 2012). From the previous researches as it is depict that there no such work on queuing system of customer care center of telecom industry in Bangladesh The authors conduct a case study on queuing system at Teletalk customer care centre, Rajshahi, Bangladesh.

3. The Methodology and Model

Data for this study is collected from Teletalk Customer Care Centre, Rajshahi. The data is collected by observation and personal interview. The following assumptions are made for queuing system at the Teletalk Customer Care Centre which is in accordance with the queue theory:

1. The queue discipline is First-In, First-Out (FIFO) basis by any of the servers. There is no priority classification for any arrival.
2. Arrivals of Customers follow a Poisson probability distribution at an average rate of λ customers per unit of time.
3. Every arrival waits to be served regardless of the length of the line; that is, there is no balking or reneging.
4. Service times also vary from one customer to the next and are independent of one another and are exponentially distributed with average μ patients per unit of time.
5. The average service rate is greater than the average arrival rate.
6. The queue can goes infinity, so there is no limit in queue.
7. Customer care officers are only considered as a server but no other personnel.

8. Service providers do not go faster because the line is longer; service rate is independent of line length.

3.1 Terminology and Notations

The following standard terminology and notation is used henceforth:

λ = mean arrival rate (expected number of arrivals per unit time)
μ = mean service rate for overall system (expected number of customer completing service per unit time by each server)
S = number of server in the queuing system
P_n = probability that exactly n customer in queuing system
Ls = expected no of customers in queuing system
Ws = expected waiting time of customers in the system
L_q = expected queue length
W_q = expected waiting time of customers in the queue
ρ = utilization factor

3.2 The M/M/S Model

The model used in this work is the (M/M/s: FIFO/∞/∞) Multi-server queuing model. This model is used in this work after studying the queuing system at Teletalk customer care centre, Rajshahi, Bangladesh. In this particular system multiple servers is required and utilized to deliver services to the customers. Service is provided according to fast in fast out method (FIFO). It is assumed that the arrivals follow a Poisson probability distribution at an average of λ customers per unit of time. It is also assumed that they are served on a first come, first-served basis by any of the servers. The service times are distributed exponentially, with an average of μ customers per unit of time and number of servers S. If there are n customers in the queuing system at any point in time, then the following two cases may arise

i. If n<S,(number of customers in the system is less than the number of servers), then there will be no queue. However, (S–n) number of servers will not be busy. The combined service rate will then be $\mu_n = \mu s$; n<s.
ii. If n≥s (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 –s). The combined service rate will be $\mu_n = \mu s$; n≥s.

Probability of having n customers in the system is given by-
The expected number of customers waiting on the queue is given by-

$$p_n = \begin{cases} \left(\rho^n/_{n!}\right) p_0 & n \le s \\ \rho^n/_{(s!\, s^{n-s})}\, p_0 & n > s;\ \rho = \lambda/_{s\mu} \end{cases}$$

$$p_o = \left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu-\lambda}\right]^{-1} \quad \text{---------- (1)}$$

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

Expected number of customers in the system-

$$L_s = L_q + \frac{\lambda}{\mu} \quad \text{----------} \quad (3)$$

Expected waiting time of customers in the queue-

$$W_q = \frac{Lq}{\lambda} \quad \text{----------} \quad (4)$$

Average time a customer spends in the queue-

$$W_s = \frac{Ls}{\lambda} \quad \text{----------} \quad (5)$$

Utilisation factor i.e, the fraction of time servers are busy-

$$\rho = \frac{\lambda}{s\mu} \quad \text{----------} \quad (6)$$

3.3 Introducing Cost Elements into the Model

To evaluate and determine the optimum number of servers in the system the following costs must be considered in making these decisions-

- **Service costs:**

They are the fairly 'tangible' costs involve in operating each service facility like the cost for equipment, materials, labor etc. These cost of course, rise as the number of service facilities put into operation increase.

- **Waiting time costs of customers:**

They are the relatively 'intangible' costs associated with causing customers to have to wait in line for some period of time prior to being waiting upon physical discomfort, adverse emotional reactions, reduced or lost sales and so on. Of course, as the number of service facilities in operation increases, the time the customer has to wait in line, on the average, decreases, and hence so too do these costs. The level of service will increase with an increase in the Level of service. So the cost will increase when we will increase level of service (number of servers) but, when we will increase the level of service expected waiting time of the customers in the system will decrease and the customers will be more satisfied. The average waiting time will also decrease with an increase in the cost of service per arrival.

3.4 Expected Total Cost

Objective Function:

$$\text{Min } \{E(TC) = E(SC) + E(WC)\}$$

Where,

E(TC): Total Expected Cost;
E(SC): Expected Cost of Providing Service;
E(WC): Expected Cost of Waiting.

The expected total cost can be represented by the following equation [2]:

$$E(TC) = S.Cs + Cw(\lambda Ws)$$

Where,

Cs: Service cost of each server
Cw: Waiting cost by customers
λ: Number of arrivals per hour
Ws: Average time an arrival spends in the system
S: Number of servers

Here it should be noted that, hourly service cost of each server is the marginal cost per hour to run a server. It is calculated as the hourly salary of each customer care officer. To determine the hourly income of the customers who are getting service from the Centre, the waiting cost of a customer is calculated by using the per capita annual income of the people of Bangladesh. This income is converted in hourly income; as a man works twenty four days in a month and eight hours in a day. In this way we have analyzed the queuing system.

4. Data Analysis and Findings

The estimated Characteristic of the System is provided in table 1. The second row of this table represents the number of servers that can be positioned at Teletalk customer care center, Rajshahi. The following data and calculations are for five servers in the system shown in third column at table 1.

- Average number of customers arriving at the system per hour, $\lambda = 35$
- The mean number of customers served per hour, $\mu = 10$
- The utilization factor for the system, $\rho = \lambda/\mu s = 35/(10*5) = .70 < 1$
- Probability that there are zero customers in the system Po = 0.0259
- The average number of customers in line waiting for service (check in) Lq =0.88162
- The average number of customers in the system Ls = 4.38162
- The average time a customer spends in the queue waiting for service Wq = 0.02519
- The average time customer spends in the waiting line or being serviced (namely, in the system) Ws = 0.12519
- The number of unoccupied booth = 5(1-.7) = 1.5 (nearly two counters are inactive)

We use TORA software to compute the performance measures of the multi-server queuing system at Teletalk Customer Care Centre, Rajshahi.

Table 1: Performance Parameters of the System at Different Number of Servers

Parameters	Number of servers						
	4	**5**	**6**	**7**	**8**	**9**	**10**
arrival rate, λ	35	35	35	35	35	35	35
service rate, μ	10	10	10	10	10	10	10
system utilization, ρ	87.5%	70%	58.33%	50%	43.75%	38.89%	35%
Lq	5.16503	0.88162	0.02845	0.0762	0.02324	0.00682	0.0019
Ls	8.66503	4.38162	3.74845	3.5762	3.52324	3.50682	3.5019
Ws, in hr	0.24757	0.12519	0.1071	0.10218	0.10066	0.10019	0.1005
Wq, in hr	0.14757	0.02519	0.0071	0.00218	0.00066	0.00019	0.00005
P_0	0.01475	0.0259	0.02896	0.02984	0.0301	0.03017	0.03019

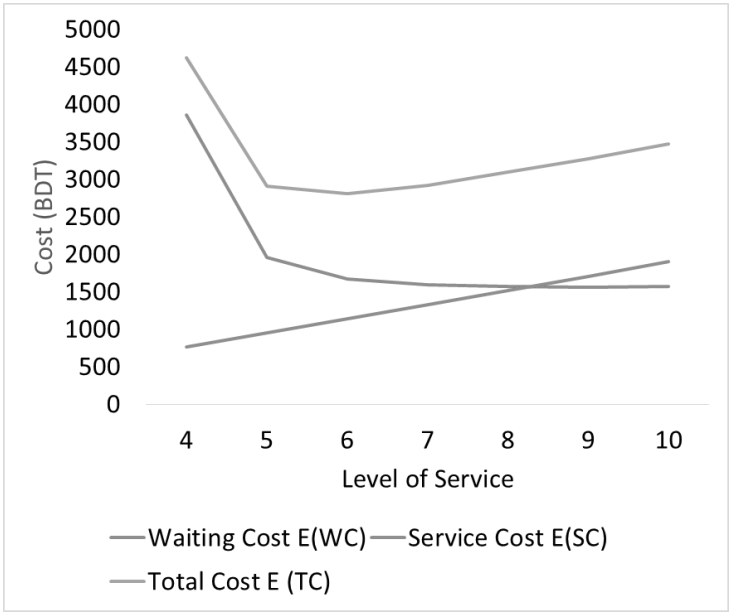
In table 2 the hourly service costs and hourly waiting costs are identified as mentioned in section 3.4 to calculate the total cost for corresponding number of servers. In this table the result shows that the minimum total cost is obtained for six servers.

Table 2: Table for Determining the Optimal Server Number at Minimum Total Cost

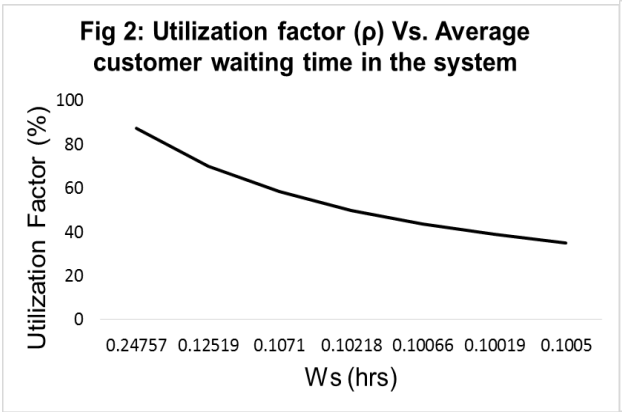
No. of servers, S	Total hourly service cost E(SC) = S.Cs	arrival rate, λ (per hour)	Expected waiting time in system Ws, in hour	Total hourly waiting cost E(WC)= Cw (λWs)	Total Expected cost E (TC) = E (SC)+ E (WC)
4	760 BDT	35	0.24757	3864.568 BDT	4624.568 BDT
5	950 BDT	35	0.12519	1954.216 BDT	2904.216 BDT
6	**1140 BDT**	**32**	**0.1071**	**1671.831 BDT**	**2811.831 BDT**
7	1330 BDT	35	0.10218	1595.03 BDT	2925.03 BDT
8	1520 BDT	35	0.10066	1571.303 BDT	3091.303 BDT
9	1710 BDT	35	0.10019	1563.966 BDT	3273.966 BDT
10	1900 BDT	35	0.1005	1568.805 BDT	3468.805 BDT

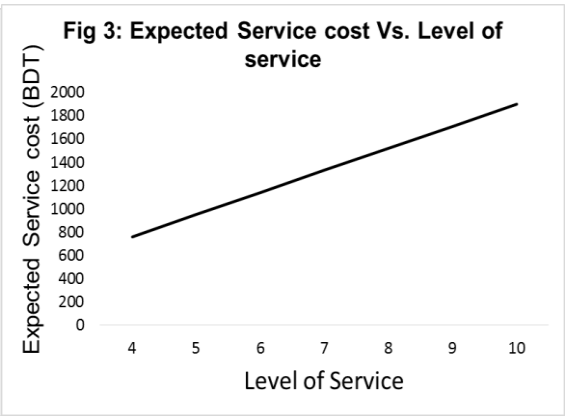
Figure 1 shows that optimal server level at the Customer Care Centre is achieved when the number of servers (Customer Care Officer) is 6 with a minimum total cost of BDT 2811.831 per hour as against the present server level of 5 Customer Care Officer at the Customer Care Centre which have high total cost of BDT 2904.216 per hour. From figure 2 it should also be noted that customers' average waiting time in the system is lower at this optimal server level compare to existing system. It is also shown from figure 4 that the Idleness of the system will also increase with an increase in number of customer in the system. Figure 3 and 5 the impact of the level of service (that is number of servers) over the service costs and the waiting costs.

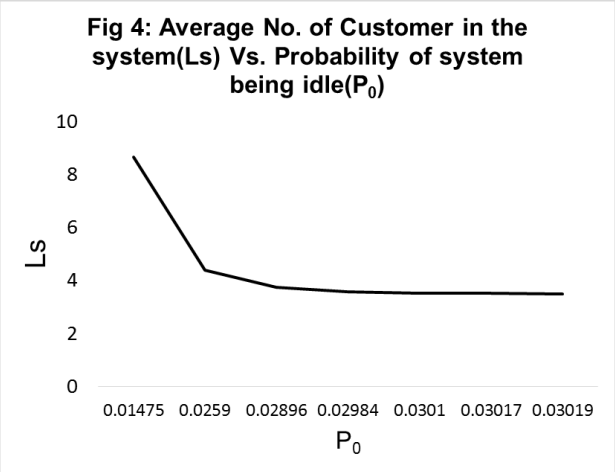
Figure 1: Chart for Determining the Optimal Server Number at Minimum Total Cost

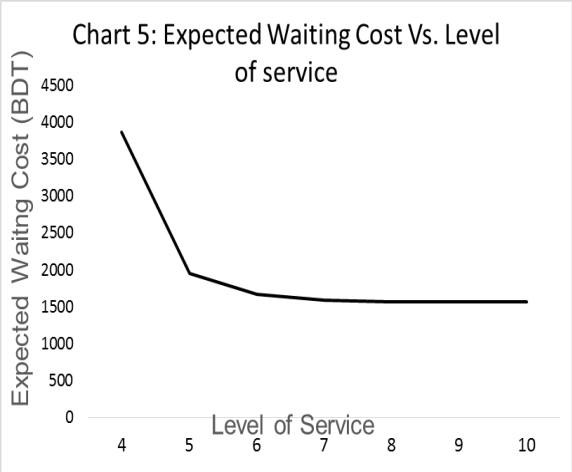


S * = 6









5. Conclusions

The queuing characteristics at the Teletalk Customer Care Centre, Rajshahi is analyzed using a Multi-server queuing Model with first in first out (FIFO) queue discipline and the Waiting and service Costs determined with a view to determine the optimal service level. The results show that average queue length, average waiting time of the customers in the system is reduced at optimum level of server from existing level of server. From the results we also see that the optimum server number is six at minimum total cost level, however the Utilization of the server is not better than the existing system. The utilization of the servers is reduced when the capacity of the customer care officer is increased from 5 to 6. In this work, the Multi-server queuing model named M/M/s: FIFO/∞/∞ is utilized to model the queuing system. However, to improve the utilization factor (ρ) of the queuing system other empirical queuing model such as M/M/s: GD/∞/∞, M/M/s: GD/N/∞, M/M/s: GD/N/N, M/M/s: FIFO/N/∞, M/M/s: FIFO/N/N etc. with considering waiting and service cost.

References

Adan, I and Resung, J 2002, 'Queuing Theory', Department of Mathematics and Computing Science, Eindhoven University of Technology, Eindhoven.

Chowdhury, MSR, Rahman, MT and Kabir, MR 2013, 'Solving of waiting lines models in the bank using queuing theory model the practice case: islami bank bangladesh limited, chawk bazar branch, chittagong', *IOSR Journal of Business and Management*, vol. 10, issue 1, pp 22-29.

Cooper, RB 1981, *Introduction to queuing theory*, 2nd edition, North Holland, New York.

Gorunescu, F, Mc Clean, SI and Millard, PH 2002, 'A queuing model for bed-occupancy management and planning of hospitals', *Journal of the Operational Research Society*, vol. 53, pp 19-24.

Gurumurthi, S and Benjaafar, S 2004, *Modeling and analysis of flexible queuing Systems*, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Hiller, FS and Liberman, GJ 2001, *Operations research*, 7th edition, McGraw Hill, New York.

Kembe, MM, Onah, ES and Lorkegh, S 2012, 'A study of waiting and service costs of a multi-server queuing model in a specialist hospital', *International Journal of Scientific & Technology Research*, vol. 1, issue 8.

Khan, MR and Callahan, BB 1993, 'Planning laboratory staffing with a queuing model', *European Journal of Operational Research*, vol. 67.

Kostas, UN 1983, *Introduction to theory of statistics*, McGraw Hill, Tokyo.

Olaniyi, TA 2004, 'An appraisal of cost of queuing in nigerian banking sector: a case study of first bank of Nigeria Plc, Ilorin', *Journal of Business & Social Sciences*, Vol.9, Nos,1&2, pp 139-145.

Panneersalvam, R 2009, *Operations research*, 2nd edition, PHI Learning, New Delhi.

Taha, HA 2007, *Operations research: an introduction*, 8th edition, Pearson, New Delhi.

Walpole, RE, Myers, RH, Myers, S and Ye, K 2009, *Probability and statistics for engineers & scientist*, 8th editionn, Pearson, New Delhi.