The Queuing Theory Application in Bank Service Optimization

Huimin Xiao, Guozheng Zhang

Institute of Information & System Engineering, Henan University of Finance & Economics, Zhengzhou 450002
E-mail: xiaohm@hnufe.edu.cn

Abstract: Lines of waiting customers are always very long in most of banks. The essence of this phenomenon is the low efficiency of queuing system. In this paper, the queuing number, the service windows number and the optimal service rate are investigated by means of the queuing theory. In technology, the optimal problem of the bank queuing is solved. The time of customer queuing is reduced. The customer satisfaction is increased. It was proved that this optimal model of the queuing is feasible. By the example, the results are effective and practical.

Key Words: The Queuing Theory, The Optimal Service Rate, $M/M/s/\infty$ Model

I. Introduction

China's commercial banks have done an great effort to increase the marketing, but the most of them are facing a serious problem- customer queuing [1], which led to low service rate of the bank the counters, poor business environment and a number of high-quality customers and potential customer lost and so on. In bank, customer queuing appears to be an unusual phenomenon in essence, it reflects the domestic commercial banks lacking of the business philosophy of customer-centric in market economy [2]. Scientists found that the in organization queuing is a major cause of the loss of customers. A survey shows that the service quality of bank branches, the waiting time of customer queuing, the convenience, and the service of the bank's staff are the best way to affect the consumer experience. The waiting time is the most impotent element in impacting on the consumer experience. In fact, the most direct complained from customers not satisfied with the bank is that the customers wait for a long time to line up.

To completely solve the problem of the long queues[2], all aspects of the bank are involved, subjecting to a variety of factors. It is not only a problem of the efficiency of business and management, but also a problem of the commercial banks innovation. In the following, the problem of customers waiting for the shortest time is studied by means of the queuing theory. the measure to reduce the time of customers queues is obtained to achieve the goal of people-oriented and the greatest effectiveness of the banks.

II. The Queuing Theory And Model Analysis

A. Queuing Theory

The queuing theory is also known as the random system theory [3,4], which studies the content of the following three parts;

(1) Behavior problems: it is that the queues probability is studied. It is mainly that the queues length distribution, the waiting time distribution and the busy period distribution are investigated in both transient state and steady-state.

(2) Optimization problem: it is divided into the static optimization and the dynamic optimization. The static optimization problem former is the optimal design, the dynamic optimization problem is the best operation of the queuing system.

(3) The statistical inference of queuing system: It is that the model of a given queuing system is determined, in order to investigate and analysis the queuing system by using the queuing theory.

B. The basic indexes of the queuing systems

State of system: the number of customers in queue systems

Queue length: the number of customers waiting for service to begin [3,5]

P_n : the probability of exactly n customers in the queue systems in the statistic equilibrium.

N : The number of the customers in the queue system in equilibrium state. The mean number/expectation of the customers is L

N_q : The number of the customers in queuing in equilibrium state. That is the expected queuing length (excludes customers being served). The mean number/expectation of the customers in a queue is L_q

T : The staying time of customers (including service time), as the queue system is in the equilibrium state; the mean number of the staying time for each individual customer in the queuing system is W

T_q : The waiting time of customers (excluding service time), as the queue system is in the equilibrium state; the mean number of the waiting time for each individual customer in the queue system is W_q .

λ_n : The mean arrival rate (expected number of arrivals per unit time) of new customers when n 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.

When λ_n is a constant for all n , it is denoted by λ .

When the mean service rate of per busy server is a constant for all $n \geq 1$, it is denoted by μ . In this case, when all s servers are

This work is supported by Henan Province Natural Science Foundation (No.072300410110)

busy, $\mu_n = s\mu$. Therefore, the expected inter-arrival time is $1/\lambda$, the expected service time is $1/\mu$. Also, $\rho = \lambda/(s\mu)$ is the utilization rate for the service facility, i.e. the expected fraction of time as the individual servers are busy, because $\lambda/(s\mu)$ represents the fraction of the system service capacity $(s\mu)$ that is being utilized on the average by arriving customers λ .

C. The $M/M/1/\infty$ Model For Single-Server Queues

This is the simplest queuing system to analyze. Here the arrival and service time are negative exponentially distributed (Poisson process). The system consists of only one server. This queuing system can be applied to a wide variety of problems as any system with a very large number of independent customers, and can be approximated as a Poisson process: $P_n = P\{N = n\}(n = 0,1,2\cdots)$ is the probability distribution of the queue length, as the queue system is in the equilibrium state, then $\lambda_n = \lambda$, $\mu_n = \mu$, $\rho = \lambda/\mu < 1$, $P_n = (1-\rho)\rho^n$, $n = 0,1,2\cdots$. The mean queue length is

$$L = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \tag{1}$$

the mean queuing length is

$$L_q = \sum_{n=0}^{\infty}(n-1)P_n = \frac{\rho^2}{1-\rho} = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{2}$$

the mean staying time is $W = E(T) = 1/(\mu-\lambda)$, the mean waiting time is $W_q = W - 1/\mu = \lambda/[\mu(\mu-\lambda)]$.

D. The $M/M/s/\infty$ Model For Multi-Server Queues

Consider an $M/M/s$ queue with arrival rate λ, service rate μ and *s* servers. The traffic intensity is defined usual by the ratio $\rho_s = \rho/s = \lambda/(s\mu)$.

When customers arrive, if the free service, you can immediately accept the services, otherwise it would line up to wait for a queue and the space is unlimited. The steady distribution of queuing system[3] is studied as following.

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

When $n < s$, the n customers of the system are receiving services, n servers are working, the efficiency rate is n times of one server, therefore, the transfer rate should be $n\mu$ from n to $n-1$.

When $n \geq s$, the number of the servers is s in the system, there are s customers who are receiving services, the efficiency rate is s times of one server, then the transfer rate should be $s\mu$, that is,

$$\mu_n = \begin{cases} n\mu & as \ \ n < s \\ s\mu & as \ \ n \geq s \end{cases} \tag{3}$$

$$P_n = \begin{cases} \dfrac{\rho^n}{n!}P_0, & n = 1,2,\cdots s\text{-}1 \\ \dfrac{\rho^n}{s!\,s^{n-s}}P_0, & n = s, s+1\cdots \end{cases} \tag{4}$$

$$P_0 = \left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!} + \frac{\rho^s}{s!(1-\frac{\rho}{s})}\right]^{-1} \tag{5}$$

The probability of n customers is given as (1) and (2), as the queuing system is under the steady state condition. When n ≥ s, it is that the number of customers of the system is not smaller than the number of servers, the next customers must wait, that is,

$$C(s,\rho) = \sum_{n=s}^{\infty} P_n = \frac{\rho^s}{s!(1-\rho_s)}P_0 \tag{6}$$

where $\rho_s = \rho/s = \lambda/s\mu$ .The probability of customers need to wait in the system is given as above (Erlang [3] waiting formula), for the multi-server systems waiting for queue, by means of the given steady distribution, we have that Mean number/ expectation L_q of participants in a queuing is

$$L_q = \sum_{n=s+1}^{\infty}(n-s)P_n = \frac{P_0\rho^s\rho_s}{s!(1-\rho_s)^2} \tag{7}$$

Mean number/expectation L of participants in a queue system is equal to that the mean queuing length L plus the mean number of customers who are receiving service, that is

$$L = L_q + \rho \tag{8}$$

For the multi-server system, Little formula is also hold, that is,

$$W = L/\lambda, \tag{9}$$

$$W_q = L_q/\lambda = W - 1/\mu \tag{10}$$

III. The Optimization In Bank Queue

In the following, by means of the queuing theory, the bank queuing problems is investigated as the following three aspects: a queue or more queues, how many service windows, what is the optimal service rate [6,7]

*A. **A Queue or Two Queues***

In reality, we have had in lining up in the bank, there are several service windows. Each service window has a queue. Because the service efficiency of the staffs is different, most of time, the customer coming after may be serviced earlier, thus, "the first come, first served" [8] principle is violated. so that customers are not satisfied. And if a team is queuing, this principle can be realized according to the order successively arrival for services. Which one is more efficient, we will analysis it from a technical point as following.

Suppose that s = 2, comparing a queue with two queues.

The customers arriving rate is λ = 32, the service rate of each platform is μ = 20, if each service platform has a queue according to their schedule, the arrival customers join in each queue at the probability 1 / 2 , this is called as the two scheduled queue. When there are two lines, the system can be regarded as two isolated M/M/1 systems, and the arrival rate of each service platform λ = λ / 2 = 16. If there is a line, the system will be M/M/2, L , L_q , W and W_q are calculated respectively, and compared

when there is a line， $s=2$， $\lambda=32$， $\mu=20$， $\rho=8/5$

$$P_0=\left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!}+\frac{\rho^s}{s!(1-\rho_s)}\right]^{-1}=\left[1+\rho+\rho^2/[2(1-\frac{\rho}{2})]\right]^{-1}$$

$$=\left[1+\frac{8}{5}+\frac{64/25}{2(1-4/5)}\right]^{-1}=\frac{1}{9}$$

$$C(s,\rho)=\frac{32}{5}\times\frac{1}{9}=\frac{32}{45}$$

$$L_q=\frac{C(s,\rho)\cdot\rho_s}{1-\rho_s}=\frac{32}{45}\times\frac{4/5}{1-4/5}=\frac{32}{45}\times 4=\frac{128}{45}=2.84$$

$$L=2.84+\frac{8}{5}=2.84+1.6=4.44$$

$$W=\frac{L}{\lambda}=\frac{4.44}{32}=0.14$$

$$W_q=\frac{L_q}{\lambda}=\frac{2.84}{32}=0.09$$

when there are two lines:

$$\lambda=16\ ,\ \mu=2\ ,\ L=\frac{\lambda}{\mu-\lambda}=\frac{16}{20-16}=4\ ,$$

$$L_q=\frac{\lambda L}{\mu}=\frac{16\times 4}{20}=3.2$$

$$W=\frac{1}{\mu-\lambda}=\frac{1}{20-16}=0.25$$

$$W_q=\frac{L_q}{\lambda}=\frac{3.2}{16}=0.2$$

when there are n lines：it is means that there are n service platforms, each service platform 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 . Mean number/expectation L of participants in a queue system is:

$$L=\frac{\lambda/n}{\mu-\lambda/n}=\frac{\lambda}{n\mu-\lambda}, \qquad (11)$$

The mean queuing length is:

$$L_q=\frac{\lambda L}{n\mu}=\frac{\lambda^2}{n\mu(n\mu-\lambda)} \qquad (12)$$

The mean staying time is:

$$W=\frac{1}{\mu-\lambda/n}=\frac{n}{n\lambda-\lambda} \qquad (13)$$

The mean waiting time is

$$W_q=\frac{nL_q}{\lambda}=\frac{\lambda}{\mu(n\mu-\lambda)} \qquad (14)$$

From the TableI, in the four main characteristics, W is the length of customers staying in the system, it is an important sign that marks the service quality of systems. In the case of the two lines, it is 0.25. At a line, it ist 0.14. The staying time is decreasing clearly, and the length of line is also less.

For a number of lines, according to the same calculating method, we give it in the table 1. This shows that in banking services, in terms of "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	L_q	W	W_q
1	32	20	4.44	2.84	0.14	0.09
2	16	20	4	3.2	0.25	0.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	$\frac{\lambda}{n}$	μ	$\frac{\lambda}{n\mu-\lambda}$	$\frac{\lambda^2}{n\mu(n\mu-\lambda)}$	$\frac{n}{n\lambda-\lambda}$	$\frac{\lambda}{\mu(n\mu-\lambda)}$

B. The Optimal Service Windows

We are now considering, in order to guarantee the quality of service we set up the number of service windows.

For customers to arrive at the rate of λ = 32, each service desk rate of μ = 20, if we need customers need to line up no more than 5% how many service windows should be set up.

when $s=2$ ， $\lambda=32$， $\mu=20$， $\rho=32/20=8/5$， $s=2$.

$$P_0=\left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!}+\frac{\rho^s}{s!(1-\rho_s)}\right]^{-1}=\left[\frac{13}{5}+\frac{64}{10}\right]^{-1}=\frac{1}{9}$$

$$C(2,\rho)=\frac{\rho^s}{s!(1-\rho_s)}P_0=\frac{64}{10}\times\frac{1}{9}=\frac{64}{90}=0.711=71.1\%$$

when $s=3$,

$$P_0=\left[1+\rho+\frac{\rho^2}{2!}+\frac{\rho^3}{3!(1-\rho/3)}\right]^{-1}$$

$$=\left[1+\frac{8}{5}+\frac{64}{50}+\frac{(8/5)^3}{6\times(1-8/15)}\right]^{-1}=\left[\frac{97}{25}+\frac{256}{7\times 25}\right]^{-1}$$

$$=\left[\frac{935}{175}\right]^{-1}=\frac{7}{37.4}$$

$$C(3,\rho)=\frac{256}{175}\times\frac{7}{37.4}=0.274=27.4\%$$

when $s=4$,

$$P_0=\left[1+\rho+\frac{\rho^2}{2!}+\frac{\rho^3}{3!}+\frac{\rho^4}{4!(1-\rho/4)}\right]^{-1}$$

$$=\left[1+\frac{8}{5}+\frac{32}{25}+\frac{(8/5)^3}{6}+\frac{(8/5)^4}{24(1-8/20)}\right]^{-1}$$

$$=\left[\frac{1711}{375}+\frac{512}{1125}\right]^{-1}=0.199$$

$$C(4,\rho)=\frac{512}{1125}\times 0.199=0.09=9\%$$

when $s=5$,

$$P_0=\left[\frac{1711}{375}+\frac{\rho^4}{4!}+\frac{\rho^5}{5!(1-\rho/5)}\right]^{-1}$$

$$=\left[\frac{1711}{375}+\frac{512}{375\times 5}+\frac{4096}{31875}\right]^{-1}=\left[\frac{9067}{375\times 5}+\frac{4096}{31875}\right]^{-1}$$
$$=0.201$$

$$C(5,\rho)=\frac{\rho^s}{s!(1-\rho_s)}\cdot P_0=\frac{4096}{31875}\times 0.201$$
$$=0.0257=2.57\%<5\%$$

We can see from the table 2, in the four main characteristics, when, the number of service windows is 5, the probability of queuing is 2.57% less then 5%. When , the used rate of service windows: $\rho_s=\rho/s=8/25=0.32=32\%$. From this bank managers could set the propriety service window to improve service.

TABLE II . MAIN CHARACTERISTICS IN THE QUEUING SYSTEMS

Number of service windows	2	3	4	5
The probability of queuing	71.1%	27.4%	19.9%	2.57%

C. The Optimal Service Rate

Here we only consider the condition of one service window, that is, consider the model *M/M/*1/∞. Take the objective function *Z* as the expectations of a unit of the cost of services in one hour and the cost of customers stay in the system. That is, $Z=c_s\mu+c_wL$.

Among them, c_s as the cost of services in one hour when $\mu=1$. c_w as the cost of customers stay in the system. We can get the following:

$$L=\lambda/(\mu-\lambda) \qquad (15)$$

So we have

$$Z=c_s\mu+c_w\cdot\lambda/(\mu-\lambda) \qquad (16)$$

From $dz/d\mu=0$, we have

$$c_s-c_w\frac{\lambda}{(\mu-\lambda)^2}=0 \qquad (17)$$

The optimal service rate is:

$$\mu^*=\lambda+\sqrt{\frac{c_w}{c_s}\lambda} \qquad (18)$$

With which we can work out the optimal services[8,9], to improve the efficiency of our services.

IV. CONCLUSION

The efficiency of commercial banks is improved by the following three measures. First, we establish the optimization model of queuing and calculate the optimal model of queuing. Second, bank should improve the business environment, and calculate the optimal number of service windows to improve operational efficiency. Third, we calculate the optimal service rate and the service efficiency by the operating costs. Of course, we should be looking for the method by which the customer waiting time and the banks cost is reduced. Queuing is essentially the problem of the domestic commercial banks lacking of the business philosophy of customer-centric in Market economy. In order to solve the queuing problem, the commercial banks should have break in the operation philosophy, change the behind concept, promote business innovation, change bank networks status in which operators mainly deal with the low-end business, raise the quality of business, reduce business costs and improve the quality of service .

REFERENCES

[1] X. X. Zhao. Queuing theory with the bank management innovation, Modern Finance, No.3, pp.9-10, 2007.

[2] R. Zhang, Analysis of the service sector queuing, Journal of Qiqihar University, No.6, pp.41-43, 2002.

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

[4] Operational research materials group, Operational Research, Beijing, China, Tsinghua University Press, 2005.

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

[6] X. G. Wang, Z. C. Jiao, Research based on queuing theory of the problem of bank, Xiangtan Normal University, Vol.30, No.1, pp.58-60, 2008.

[7] C. H Zhu, Queuing theory based on the parallel computer system for analysis of the evaluation, Heilongjiang Engineering University, Vol.16, No.2, pp.53-55, 2002.

[8] J. Li, Queuing theory and the witness service used in supermarkets of the optimal design, Chinese Information Technology Management, vol.11, No.13, pp.75-78, 2008.

[9] H. Zeng, Queue based on the cost of distribution systems, Journal of Wuhan University of Science and Technology, Vol.30, No.3, pp330-332, 2007.