

# Optimal Decision of M/M/C(m) Queuing System

Xiaomin Zhao

School of Management, Shanghai  
University,  
Shanghai, China

Junjun Gao

Sydney Institute of Commerce,  
Shanghai University, Shanghai, China

Yihua Wei

School of Management, Shanghai  
University,  
Shanghai, China

**Abstract**—To study the M/M/C(n) queuing system in which a sever can service multiple customers at the same time, the stability equations is constructed, by which the stability probability is solved, and then some system operation indexes are obtained. Furthermore, based on those operation indexes, the number of customers on simultaneous service is optimized, some properties for the optimal solutions are obtained, and then two algorithms are given to solve the optimal problem.

**Keywords**—queuing system; simultaneous service; operation index; algorithms

## I. Introduction

Queuing theory have been widely used in so many areas, such communication, computer net work, business, etc. In traditional queuing models, one server can only service one customer at the same time. But in many practice, for example, a server can service multiple customers simultaneously. For example, a network server can service to multiple clients simultaneously, server of an online customer service system can provide customer service for multiple users, etc.

When a server can service multi-customer at a same time, the system's operation efficiency will be effect by the number of customers on simultaneous service. For different service number, how does the queuing system operation efficiency change? Which service number can optimize the system efficiency? We will study those problems.

The remainder of the paper is organized as follows. Section 2 describes our modeling assumptions and constructs the system model. Section 3 discusses the optimize models, and some properties for the optimal solutions are obtained, based on which algorithms are given to solve optimal problems. Section 4 gives a numerical example to illustrate our results. Finally, Section 5 is a concluding section.

## II. Model

Consider the following M/M/C(m) queuing system: Customers arrive according to a homogeneous Poisson process with intensity  $\lambda$ . The number of server is  $C$ , and each server can service multi-customer simultaneously. The service time of each service is an independent negative exponential distribution random variable. The service rate of server  $i, i = 1, 2, \dots, C$  is  $\mu_i(m)$ , which depend on the number of customer on simultaneous service,  $m$ .

Given the number of customer on simultaneous service for server  $i, i = 1, 2, \dots, C$  by  $m_i$ , where  $\sum_{i=1}^C m_i = k$ , the service rate of the queuing system will be

$$\mu_k(m_i) = \sum_{i=1}^C m_i / t(m_i). \quad (1)$$

The maximum service rate of the queuing system with  $k$  on service customer is  $u_k = \max_{m_i} u_k(m_i)$ .

The decision maker can adjust the queuing system's simultaneous service number  $n$ . If the customer number  $k$  no more than  $n$ , the queuing system can simultaneous service all  $k$  customers, and the service rate is  $\mu_k$ ; else, the queuing system can only simultaneous service  $n$ , and the service rate is  $\mu_n$ .

Denote  $P_k$  represent the stability probability of that the re are  $k$  customer in the queuing system. By use of the related theory of markov chain, we can obtain the following state transition equation:

$$\begin{cases} -\lambda P_0 + \mu_1 P_1 = 0, \\ \lambda P_{k-1} + \mu_{k+1} P_{k+1} - (\lambda + \mu_k) P_k = 0, & k = 1, 2, \dots, n-1 \\ \lambda P_{k-1} + \mu_n P_{k+1} - (\lambda + \mu_n) P_k = 0, & k = n, n+1, \dots \end{cases} \quad (2)$$

Rearrange the above equation, we can obtain:

$$P_k = \begin{cases} \prod_{i=1}^k \rho_i P_0, & k = 1, 2, \dots, n \\ (\rho_n)^{k-n} \prod_{i=1}^n \rho_i P_0, & k = n+1, n+2, \dots \end{cases} \quad (3)$$

where  $\rho_k = \lambda / \mu_k$  is called traffic intensity.

Also because that  $\sum_{k=0}^{\infty} P_k = 1$ , so if  $\rho_n < 1$ , we can obtain

$$P_0 = \frac{1}{1 + \sum_{k=1}^{n-1} \prod_{i=1}^k \rho_i + \frac{1}{1-\rho_n} \prod_{i=1}^n \rho_i} \quad (4)$$

Substituting this into formula (3), we can obtain the solution of  $P_k$ .

If  $\rho_n \geq 1$ , then the customer number approach infinity with probability 1, the queuing system will never approach stability.

By use of the solution of  $P_k$ , we can obtain the follows system stability operating indexes:

1) Average Waiting Queue Length (average waiting customer number in the system)

$$\begin{aligned} L_s(n) &= \sum_{k=n}^{\infty} (k-n)P_k \\ &= \sum_{k=n}^{\infty} (k-n)(\rho_n)^{k-n} \prod_{i=1}^n \rho_i P_0 \\ &= \prod_{i=1}^n \rho_i P_0 \sum_{l=0}^{\infty} l(\rho_n)^l \\ &= \frac{\rho_n \prod_{i=1}^n \rho_i}{(1-\rho_n)^2} P_0 \end{aligned} \quad (5)$$

2) Average Queue Length (average customer number in the system)

$$\begin{aligned} L_q(n) &= \sum_{k=0}^{\infty} kP_k \\ &= \sum_{k=1}^{n-1} k \prod_{i=1}^k \rho_i P_0 + \sum_{k=n}^{\infty} n(\rho_n)^{k-n} P_n + L_s \\ &= \left[ \sum_{k=1}^{n-1} k \prod_{i=1}^k \rho_i + \frac{n-(n-1)\rho_n}{(1-\rho_n)^2} \prod_{i=1}^n \rho_i \right] P_0 \end{aligned} \quad (6)$$

3) Average Waiting Time of Customers

$$W_s(n) = L_s(n) / \lambda \quad (7)$$

4) Average Stay Time of Customers

$$W_q(n) = L_q(n) / \lambda \quad (8)$$

Smaller the value of the above four indexes, better the operation situation of the queuing system. In the next section, we will discuss how to control the number of simultaneous serviced customer of the queuing system to minimize those indexes.

### III. Decision of the Number of Simultaneous Service Customer $n$

In the four operation index, Average Waiting Queue Length and Average Waiting Time of Customers are consistently, Average Queue Length and Average Stay Time of Customers are consistently. Therefore, we only need to discuss the following two problems.

1) Minimize the Average Waiting Queue Length

$$\min L_s(n) = \frac{\rho_n \prod_{i=1}^n \rho_i}{(1-\rho_n)^2} P_0 \quad (9)$$

Denote the optimal solution of which by  $N_s$ .

2) Minimize the Average Queue Length

$$\max L_q(n) = \left[ \sum_{k=1}^{n-1} k \prod_{i=1}^k \rho_i + \frac{n-(n-1)\rho_n}{(1-\rho_n)^2} \prod_{i=1}^n \rho_i \right] P_0 \quad (10)$$

Denote the optimal solution of which by  $N_q$ .

Let  $N_1 = \min\{l \mid \mu_l = \max \mu_k, k=1,2,\dots\}$  and  $N_2 = \max\{l \mid \mu_l = \max \mu_k, k=1,2,\dots\}$  denote the maximum and minimum customer number that maximize or minimize the service rate  $\mu_k$  respectively. Obviously,  $N_1 \leq N_2$ . Let  $N_3$  denote the maximal  $k$  that makes service rate  $\mu_k$  grater than arrive rate  $\lambda$ .

Follow from the forgoing discussing, we know that only if  $\rho_n < 1$ , the queuing system can approach to stability. Because that for any given  $n > N_3$ , the system's traffic intensity no less than 1, so we only need to discuss model (9) and (10) in the in the bound of  $\{1, 2, \dots, N_3\}$ .

The optimal solution  $N_s$  and  $N_q$  have the following property.

**Theorem 1**  $N_q \leq N_1 \leq N_2 \leq N_s \leq N_3$ . In the other word, the optimal solution of model (9) no less than  $N_2$ , and the optimal solution of model (10) no more than  $N_1$ .

**Prof.** 1) We will prove that  $N_2 \leq N_s$  firstly. We only need to prove that for all  $n < N_2$ ,  $W_s(n) > W_s(N_2)$ , or  $L_s(n) > L_s(N_2)$ .

$$\begin{aligned} L_s(n) &= \frac{\rho_n \prod_{i=1}^n \rho_i}{(1-\rho_n)^2} P_0 \\ &= \frac{\rho_n \prod_{i=1}^n \rho_i}{(1-\rho_n)^2} \frac{1}{1 + \sum_{k=1}^{n-1} \prod_{i=1}^k \rho_i + \frac{1}{1-\rho_n} \prod_{i=1}^n \rho_i} \\ &= \frac{\rho_n}{1-\rho_n} \frac{1}{1 + (1-\rho_n) \sum_{k=0}^{n-1} \prod_{i=k+1}^n \frac{1}{\rho_i}} \\ &> \frac{\rho_n}{1-\rho_n} \frac{1}{1 + (1-\rho_n) \sum_{k=0}^{N_2-1} \prod_{i=k+1}^{N_2} \frac{1}{\rho_i}} \\ &\geq \frac{\rho_{N_2}}{1-\rho_{N_2}} \frac{1}{1 + (1-\rho_{N_2}) \sum_{k=0}^{N_2-1} \prod_{i=k+1}^{N_2} \frac{1}{\rho_i}} \\ &= L_s(N_2), \end{aligned}$$

Where the first inequality comes from that  $\sum_{k=0}^{n-1} \prod_{i=k+1}^n \frac{1}{\rho_i}$  is increasing in  $n$ , the second inequality comes from  $\rho_n \geq \rho_{N_2}$ .

2) We next prove  $N_q \leq N_1$ . We only need to prove that for all  $n > N_1$ ,  $W_q(n) > W_q(N_1)$  or  $L_q(n) > L_q(N_1)$ .

For any  $n > N_1$ , follow from the definition of  $N_1$ , we have  $\rho_n \leq \rho_{N_1}$ , so that  $L_q(n) > L_q(N_1)$ .

Combine 1) and 2), the theorem is proved.  $\square$

For Convenient, denote

$$A_n = \sum_{k=0}^{n-1} \prod_{i=k+1}^n \frac{1}{\rho_i}, B_n = \sum_{k=1}^{n-1} k \prod_{i=k+1}^n \frac{1}{\rho_i}.$$

Then formula (5) and (6) can be rewritten to

$$L_s(n) = \frac{\rho_n}{(1-\rho_n)^2 A_n + (1-\rho_n)}$$

$$L_q(n) = \frac{(1-\rho_n)^2 B_n + n - (n-1)\rho_n}{(1-\rho_n)^2 A_n + (1-\rho_n)} \quad (11)$$

$A_n$  and  $B_n$  have the following recurrence relations

$$A_1 = \frac{1}{\rho_1}, B_1 = 0$$

$$A_{n+1} = \sum_{k=0}^n \prod_{i=k+1}^{n+1} \frac{1}{\rho_i} = \frac{1}{\rho_{n+1}} \sum_{k=0}^{n-1} \prod_{i=k+1}^n \frac{1}{\rho_i} + \frac{1}{\rho_{n+1}} = \frac{1+A_n}{\rho_{n+1}}$$

$$B_{n+1} = \sum_{k=1}^n k \prod_{i=k+1}^{n+1} \frac{1}{\rho_i} = \frac{1}{\rho_{n+1}} \sum_{k=1}^{n-1} k \prod_{i=k+1}^n \frac{1}{\rho_i} + \frac{n}{\rho_{n+1}} = \frac{n+B_n}{\rho_{n+1}} \quad (12)$$

By use of Theorem 1 and recurrence rations formula (12), we can give follows algorithms to solve optimal model (9) and (10) respectively:

**Algorithm 1 (Minimize the Average Waiting Queue Length/ Average Waiting Time of Customers)**

**Step1:** Input value of each parameter, and let  $n = N_2$ ,

$$A_{N_2} = \sum_{k=0}^{N_2-1} \prod_{i=k+1}^{N_2} \frac{1}{\rho_i}, L_s = \infty, N_s = 0;$$

**Step2:** If  $\rho_n < 1$ , then compute  $L_s(n)$  by formula (11);

**Step3:** If  $L_s(n) < L_s$ , then  $L_s = L_s(n)$ ,  $N_s = n$ ;

**Step4:** If  $n < N_3$ , then  $A_{n+1} = \frac{1+A_n}{\rho_{n+1}}$ ,  $n = n+1$ , go to step 2;

**Step5:** Output the minimal average waiting queue length  $L_s$ , the minimal average waiting time of customers  $W_s = L_s / \lambda$ , and the related optimal number of simultaneous serviced customer  $N_s$ .

**Algorithm 2 (Minimize the Average Queue Length/ Average Stay Time of Customers)**

**Step1:** Input value of each parameter, and let  $n = 1$ ,  $A_1 = 1/\rho_1$ ,  $B_1 = 0$ ,  $L_q = \infty$ ,  $N_q = 0$ ;

**Step2:** If  $\rho_n < 1$ , then compute  $L_s(n)$  by formula (11);

**Step3:** If  $L_q(n) < L_q$ , then  $L_q = L_q(n)$ ,  $N_q = n$ ;

**Step4:** If  $n < N_1$ , then  $A_{n+1} = \frac{1+A_n}{\rho_{n+1}}$ ,  $B_{n+1} = \frac{n+B_n}{\rho_{n+1}}$ ,  $n = n+1$ , go to step 2;

**Step5:** Output the minimal average queue length  $L_q$ , the minimal average stay time of customers  $W_q = L_q / \lambda$ , and the related optimal number of simultaneous serviced customer  $N_q$ .

#### IV. Numeral Study

Consider a queuing system, there is 1 server, customers arriving according to a Poisson process with intensity 2; the service rate is

$$\mu_k = \frac{16k}{12+k(k-3)},$$

And the related traffic intensity is

$$\rho_k = \frac{\lambda}{\mu_k} = \frac{12+k(k-3)}{8k}.$$

By analysis  $\mu_k$ , when  $k=3$  or 4,  $\mu_k$  approach its maximum 0.5, thus  $N_1=3$  and  $N_2=4$ ; when  $k \geq 10$ ,  $\rho_k > 1$ , thus  $N_3=9$ . Therefore, we only need to discuss the optimal solution from 1 to 9.

Table 1 gives operation indexes of the system for different  $n$ . In which, when  $n=1$ , traffic intensity  $\rho_n=1.25>1$ , the system can not approach stability; when  $n=3=N_1$ , the average queue length obtains minimum 2.0902, and the average stay time obtains minimum 1.0451; when  $n=5>N_2$ , the average waiting queue length obtains minimum 0.0757, and the average waiting time obtains minimum 0.0378.

Table 1 Operation indexes for different  $n$

$n$	$\rho_n$	$L_s(n)$	$W_s(n)$	$L_q(n)$	$W_q(n)$
1	1.2500	-	-	-	-
2	0.6250	0.8013	0.4006	2.3397	1.1699
3	<b>0.5000</b>	0.2049	0.1025	<b>2.0902</b>	<b>1.0451</b>
4	<b>0.5000</b>	0.1025	0.0512	2.1926	1.0963

5	0.5500	<b>0.0757</b>	<b>0.0378</b>	2.2999	1.1499
6	0.6250	0.0764	0.0382	2.4080	1.2040
7	0.7143	0.1060	0.0530	2.5487	1.2744
8	0.8125	0.2225	0.1112	2.8274	1.4137
9	0.9167	1.0946	0.5473	4.1334	2.0667

## V. Conclusion

We studied a queueing system in which a server can service multi-customer simultaneously. For given simultaneous service number, we obtained four operations indexes of the system: Average Waiting Queue Length, Average Queue Length, Average Waiting Time, and Average Stay Time. Based those indexes, optimize models

are constructed, and some properties for the optimal solutions are obtained, and then two algorithms are given to solve the two optimal problems respectively.

## References

- [1] L. Kleinrock, Queueing Systems, vol.I, Wiley, 1974.
- [2] L. Breuer and D. Baum. An introduction to queueing theory and matrix-analytic methods. Springer, 2005.
- [3] Ivo J. B. F. Adan, Onno J. Boxma, Jacques Resing. Queueing Models with Multiple Waiting Lines. Queueing System, 2001, No. 3, pp. 65-98.
- [4] Martin J. Beckmann. Making the customers wait: The optimal number of servers in a queueing system. OR Spectrum, 1994, No.2, pp. 77~79.
- [5] Tao Yang, Wuyi Yue, Jianfen Zhan, et al. Optimisation of a generalised M(k)/M/k queueing system. International Journal of Revenue Management, 2009, No. 4, 428-440.