Study on the queuing system for services being processed parallel by multi-servers

WANG Guang[1,a], MA Qiaozheng[2,b], LI Xiangjun[1,c]

[1]School of Information Engineering, Xi'an University, Xi'an, 710065, China;

[2]School of Foreign Studies, Xi'an University, Xi'an, 710065, China

[a]email: wangguang@xawl.edu.cn, [b]email:maqz@163.com, [c]email: leelindass@163.com

Keywords: Queuing System; Parallel Service; Parallel Capacity; Running Index

Abstract. In view of the fact that a server can provide parallel service for multiple customers in the practical application at the same time, the corresponding queuing models are constructed, and the system running indexes are inferred accordingly such as the steady-state probability, average queue length, average waiting length, average staying time and average waiting time. Further, the running indexes are optimized with the parallel capacity as one of the parameters, the relevant optimization models are established simultaneously. Finally, by demonstrating some properties of the optimal strategy in the queuing system, the algorithm of solving optimization model is proposed. Example study result shows that the control of the parallel capacity can effectively optimize the system running indexes, especially for the average waiting length and the average waiting time.

Introduction

The queuing study can be divided into two categories: one is the application of the classical theory of queuing model[1] into the specific situations, such as solving the cloud computing problems[2], applying in the manufacturing industry[3], dealing with the problems of changes in demand[4]; the other is the theoretical study of the ever-rising queuing problems in practice based on the study of classical queuing system. Theoretical study of vacation server is conducted in literature [5-7]. The queuing problem based on a repairable server is also studied in literature[8-9]. The queuing problem of the relevance between the customer arrival rate and the number of servers is studied, based on the classical model in M/M/k[10].

In the traditional queuing model, each server can serve only one customer at a time. But now, there are new queuing problems to be solved, for example, a server has to serve several customers simultaneously, which is just like a CPU can process multiple processes in parallel at the same time, or one clerk at the Taobao online customer service center can serve multiple clients at the same time, and so on. For this kind of queuing problems, what is the operational condition of the queuing system? How can the number of parallel services be controlled, can we optimize the indicators of the queuing system? Based on the classical M/M/k model, this paper is aimed to study the queuing problems of multi-servers serving several customers in parallel simultaneously.

Model Assumption

Consider the queuing system as follows: customers' arrival follows the Poisson process with the parameter λ; suppose the servers in the queuing system is C, each server can provide services for multiple customers in parallel, the servicing time follows the independent negative exponential distribution, the average service time is $t(m_i)$ for the service desk *i (i=1,…,C)* to serve mi customers.

The average number of customers, served by the ith server in a time of unit, is its service rate $m_i/t(m_i)$. At this moment, the number of customers $k=\sum_{i=1}^{C} m_i$, served by the entire system in parallel, the corresponding service rate is $\sum_{i=1}^{C} m_i/t(m_i)$. The maximum service rate μ_k when the system accommodates k customers in parallel can be obtained by the following equation:

$$\mu_k = \max_{m_i, i=1,\dots,C} \sum_{i=1}^{C} m_i / t(m_i)$$

$$s.t. \qquad \sum_{i=1}^{C} m_i = k \tag{1}$$

In general, when the number of customers, served by the system in parallel is small, the resource of the servers is not made full use of. With the increase of the customer number k in parallel service, the utilization rate of the system servers increase gradually, and the same does the system service rate μ_k; when the number of customers in parallel service increases to an upper bound or a threshold, the utilization rate of the server reaches its maximum capacity; the server's utilization rate can no longer be increased only by increasing the number k of customers in parallel services; instead, it will take more time for the server to switch between different customers, making the service rate μ_k decline steadily.

Hence, it doesn't necessarily mean that the more customers served at the same time, the better the service rate obtained. Here we define the system parallel ability as the maximum number of customers parallelly served at the same time, noted as n。When the number of customers k is less than n, the system can serve k customers at the same time, and the system service rate is μ_k; when the number of customers in the system is larger than n, the system can serve for n customers parallelly, and the system service rate is μ_n.

Running Indexes of the Queuing System

Let $P_k(k = 1, \dots, n)$ denote the probability when there are k customers in the system under the steady condition, by using the relevant knowledge of Markov chains, it can meet the state transition equation as follows:

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

The above recursive equation can be transformed as below:

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

where $\rho_k = \lambda / \mu_k$ refers to the service intensity.

Since $\sum_{k=0}^{\infty} P_k = 1$, when $\rho_n < 1$, solving equation (3), we can the value of p_0 from equation (4) expressed as follows:

$$P_0 = \frac{1}{1 + \sum_{k=1}^{n-1} \prod_{i=1}^{k} \rho_i + \frac{1}{1-\rho_n} \prod_{i=1}^{n} \rho_i}, \tag{4}$$

and then, substituting p_0 into equation (3), we can obtain the probability P_k of k customers in the queuing system.

If $\rho_n \geq 1$, that means the number of customers in the system tends to infinite with the probability value equals to 1, the system will become congested, break down, and never reach to the steady state. It should be noted that the precondition for the system running steadily is that the service intensity ρ_n is less than 1 when the server can serve n customers parallelly; and when the number of parallelly served customers is less than n, it doesn't matter whether the service intensity ρ_n is less than 1 or not. For instance, in the case of n=2, λ =2, μ1=1, μ2=4, where $\rho_1 = \lambda/\mu_1 = 2 > 1$, $\rho_2 = \lambda/\mu_2 = 0.5 < 1$, the system still runs steadily, and its probability of stability is $P_0 = 0.2$, $P_k = 0.2 \times 2^{2-k}$, $k = 1, 2, \dots$.

The following four running indexes for the queuing system are generally taken into

consideration:

1) Average waiting queue length (the number of waiting customers in the queuing system);

$$L_s(n) = \sum_{k=n}^{\infty}(k-n)P_k$$

2) Average queue length (the number of customers in the queuing system) ;

$$L_q(n) = \sum_{k=0}^{\infty} kP_k$$

3) Average waiting time for customers $W_s(n) = L_s(n)/\lambda$;

4) Average staying time for customers $W_q(n) = L_q(n)/\lambda$.

The smaller the four index values, the better the performance of the system. Obviously, among these four indexes, the two indexes for the average waiting queue length and the average waiting time are consistent, while the other two indexes for the average queue length and the average staying time are consistent. There are only two indexes, the average waiting queue length and the average queue length, to be further discussed.

Theorem 1: The analytical expressions for the average waiting queue length and the average queue length are shown in equation (5) respectively.

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

$$L_q(n) = \left[\sum_{k=1}^{n-1} k \prod_{i=1}^{k} \rho_i + \frac{n-(n-1)\rho_n}{(1-\rho_n)^2} \prod_{i=1}^{n} \rho_i\right] P_0 \tag{5}$$

Proof. Substituting equation (3) into the expression for the average waiting queue length, we can obtain that

$$L_s(n) = \sum_{k=n}^{\infty}(k-n)P_k = \sum_{k=n}^{\infty}(k-n)(\rho_n)^{k-n}\prod_{i=1}^{n}\rho_i P_0$$

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

$$= \frac{\prod_{i=1}^{n}\rho_i P_0}{1-\rho_n}\left(\sum_{k=1}^{\infty}k(\rho_n)^k - \sum_{k=2}^{\infty}(k-1)(\rho_n)^k\right) = \frac{\prod_{i=1}^{n}\rho_i P_0}{1-\rho_n}\sum_{k=1}^{\infty}(\rho_n)^k$$

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

and meanwhile substituting equation (3) into the expression for the average queue length, we can obtain that

$$L_q(n) = \sum_{k=0}^{\infty}kP_k = \left[\sum_{k=1}^{n-1}k\prod_{i=1}^{k}\rho_i + \sum_{k=n}^{\infty}k(\rho_n)^{k-n}\prod_{i=1}^{n}\rho_i\right]P_0$$

$$= \left[\sum_{k=1}^{n-1}k\prod_{i=1}^{k}\rho_i + n\prod_{i=1}^{n}\rho_i\sum_{k=n}^{\infty}(\rho_n)^{k-n} + \sum_{k=n}^{\infty}(k-n)(\rho_n)^{k-n}\prod_{i=1}^{n}\rho_i\right]P_0$$

$$= \left[\sum_{k=1}^{n-1}k\prod_{i=1}^{k}\rho_i + n\prod_{i=1}^{n}\rho_i\sum_{k=0}^{\infty}(\rho_n)^k\right]P_0 + L_s(n)$$

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

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

Q.E.D □

Theorem 1 shows the analytical expressions for the average waiting queue length and the average queue length, from which we can learn that the two values of $L_s(n)$ and $L_q(n)$ are related to the parallel capacity n.

For the sake of convenience, we denote A_n and B_n as below:

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

Then, equation (5) can be expressed as

$$L_s(n) = \frac{\rho_n}{(1-\rho_n)^2 A_n + (1-\rho_n)}, \quad L_q(n) = \frac{(1-\rho_n)^2 B_n + n - (n-1)\rho_n}{(1-\rho_n)^2 A_n + (1-\rho_n)} \tag{6}$$

A_n and B_n have the following recurrence relation

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

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

$$B_{n+1} = \sum_{k=1}^{n} k \prod_{i=k+1}^{n+1} \frac{1}{\rho_i} = \frac{1}{\rho_{n+1}} \sum_{k=1}^{n-1} k \prod_{i=k+1}^{n} \frac{1}{\rho_i} + \frac{n}{\rho_{n+1}} = \frac{n+B_n}{\rho_{n+1}} \tag{7}$$

Decision Making Based on Parallel Capacity

We will discuss how to control the parallel capacity n so as to optimize the two running indexes --- the average waiting queue length and the average queue length.

1) The decision model for the shortest average waiting queue length is

$$\min_{n=1,2,\ldots} L_s(n) = \frac{\rho_n \prod_{i=1}^{n} \rho_i}{(1-\rho_n)^2} P_0 \tag{8}$$

Denote its optimal solution as N_s.

2) The decision model for the shortest average queue length is

$$\max_{n=1,2\ldots} L_q(n) = \left[\sum_{k=1}^{n-1} k \prod_{i=1}^{k} \rho_i + \frac{n-(n-1)\rho_n}{(1-\rho_n)^2} \prod_{i=1}^{n} \rho_i \right] P_0 \tag{9}$$

Denote its optimal solution as N_q.

Let M_1 represents the set of all k_s in which $\mu_k > \lambda$ (μ_k: the service rate, λ: the arrival rate), namely, the service intensity $\rho_k < 1$, where the maximum value is denoted as N_3. It is obvious that the value for the optimal parallel capacity n needs to be discussed within the scope of M_1. Let M_2 represents the set of the maximum points of the service rate μ_k, the minimum and the maximum values are regarded as N_1,N_2 respectively. If the maximum value μ_k is unique, then $N_1 = N_2$ 。

The optimal solution for Model (8) has the following properties:

Theorem 2: The value of the optimal parallel capacity N_s for the average waiting queue length (the average waiting time) is not less than the peak value N_2 of the service rate μ_k, say $N_s \geq N_2$.

Proof. Here what we need to prove is that for any $n \leqslant N_2$, the inequalities $W_s(n) > W_s(N_2)$ or $L_s(n) > L_s(N_2)$ hold.

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

$$= \frac{\rho_n \prod_{i=1}^{n} \rho_i}{(1-\rho_n)^2} \frac{1}{1 + \sum_{k=1}^{n-1} \prod_{i=1}^{k} \rho_i + \frac{1}{1-\rho_n} \prod_{i=1}^{n} \rho_i}$$

$$= \frac{\rho_n}{1-\rho_n} \frac{1}{1 + (1-\rho_n) \sum_{k=0}^{n-1} \prod_{i=k+1}^{n} \frac{1}{\rho_i}}$$

$$> \frac{\rho_n}{1-\rho_n} \frac{1}{1 + (1-\rho_n) \sum_{k=0}^{N_2-1} \prod_{i=k+1}^{N_2} \frac{1}{\rho_i}}$$

$$\geq \frac{\rho_{N_2}}{1-\rho_{N_2}} \frac{1}{1+(1-\rho_{N_2})\sum_{k=0}^{N_2-1}\prod_{i=k+1}^{N_2}\frac{1}{\rho_i}}$$
$$= L_s(N_2),$$

The first inequality from $\sum_{k=0}^{n-1}\prod_{i=k+1}^{n}\frac{1}{\rho_i}$ will increase progressively with the increase of n; the second inequality comes from $\rho_n \geq \rho_{N_2}$. Q.E. D □

The optimal solution Nq for Model (9) can be obtained from the following theorem:

Theorem 3: The peak value N_1 of the service rate μ_k can make the average queue length (average time) to get its optimal value, namely, $N_q = N_1$.

Proof. As for the queuing system with any parallel capacity $n \neq N_1$, the customer arrival rate, in any case, is similar to that with parallel capacity N1, yet the service rate is less than the latter, so the number of the average staying customers is larger than the latter, say, $Lq(n) \geqslant Lq(N1)$. Q.E. D □

The algorithm for solving model (8) for the solution to the optimal queue length can be obtained by applying Theorem 2 and Equations (6) and (7):

Algorithm 1 Calculation of the Optimal Queue Length/Time

s1：Input each parameter, let $n=N_2$, $A_{N_2} = \sum_{k=0}^{N_2-1}\prod_{i=k+1}^{N_2}\frac{1}{\rho_i}$, $L_s = \infty$, $N_s = 0$；

s2：If $\rho_n < 1$, calculate $L_s(n)$ by applying equation (11)；

s3：If $L_s(n) < L_s$, then let $L_s = L_s(n)$, $N_s = n$；

s4：If $n<N_3$, then let $A_{n+1} = \frac{1+A_n}{\rho_{n+1}}$, $n = n+1$, go to s2；

s5：Output the optimal waiting queue length L_s, the optimal waiting time $W_s = L_s/\lambda$, and the corresponding parallel capacity N_s.

Example Study

Consider a queuing system, where customers follow the Poisson Process, there will be two customers arrived every hour, namely, the arrival rate is λ=2；there is only one server in the queuing system, the service time follows negative index distribution. When the server serves m customers at the same time, the average service time is:

$$t(m) = \begin{cases} \frac{3}{4} + \frac{1}{16}m(m-3) & m = 1,2,3 \\ \frac{15}{17} + \frac{5}{136}m(m-3) & m = 4,5,6,... \end{cases}$$

Therefore, the service rate of the queuing system can be represented as follows:

$$\mu_k = \begin{cases} \dfrac{16k}{12+k(k-3)} & k = 1,2,3 \\ \dfrac{36k}{120+5k(k-8)} & m = 4,5,6,... \end{cases},$$

From the above we can get the set $M_1(M_1=\{2,3,4,5,6,7,8\})$ in which the service rate μ_k is larger than the arrival rate λ $(\lambda=2)$, where the set of maximum value of μ_k is $M_1=\{3, 5\}$, and μ_k max is 0.5. Hence, $N_1=3$、$N_2=5$. Based on the Theorems 2 and 3, it can be obtained that the optimal solution for the shortest waiting queue length (the shortest waiting time) must be in the set of {5,6,7,8}, and the optimal solution for the shortest queue length (the shortest staying time) is 3.

Table 1 shows each running index under different values for the parallel capacity n. When value n is not in the set of M_1, the system will break down, so it is not necessary to take it into account. When $n=3=N_1$, the queue length $L_q(n)$ reaches to the minimum vale 1.5574 and the staying time $W_q(n)$ to 0.7787 hours; when $n=6>N_2$, the waiting queue length $L_s(n)$ reaches to its minimum value 0.0437, and the waiting time $W_s(n)$ to 0.0218 hours.

Besides, table 1 also shows that the parallel capacity n has a greater effect on the waiting queue length (waiting time) than on the average queue length (staying time). The average shortest waiting

time is 0.0218 hours, the longest is 0.4006 hours, with the difference of nearly 20 times; while the average shortest staying time is 0.7787 hours, the longest is 1.0256 hours, with the difference of only a little over 30 %.

Table 1 Running Indexes of the System under Different Parallel Capacity n

n	Service rate μ_n	Service intensity ρ_n	Average waiting queue length $L_s(n)$	Average waiting time $W_s(n)$	Average queue length $L_q(n)$	Average staying time $W_q(n)$
2	3.2000	0.6250	0.8013	0.4006	2.0513	1.0256
3	**4.0000**	**0.5000**	0.2049	0.1025	**1.5574**	**0.7787**
4	3.6000	0.5556	0.1561	0.0780	1.6746	0.8373
5	**4.0000**	**0.5000**	0.0563	0.0281	1.5961	0.7981
6	3.6000	0.5556	**0.0437**	**0.0218**	1.6426	0.8213
7	2.9647	0.6746	0.0659	0.0329	1.7521	0.8761
8	2.4000	0.8333	0.2509	0.1255	2.1435	1.0717

Conclusion

This paper studies the queuing system that a server can serve multiple customers at the same time. Given the parallel capacity, the corresponding queuing models are built, the analytical expressions for the running indexes are inferred such as the probability of stability, the average queue length, the average waiting queue length, the average staying time, the average waiting time. Moreover, the system running indexes are optimized through the control of the parallel capacity n. Furthermore, by demonstrating several properties of the optimal strategies in a queuing system, the algorithm for solving the optimization model is proposed, offering the theoretical support for the practical applications. Example study result shows that the control of the parallel capacity can effectively optimize the system running indexes such as the average queue length, the average waiting queue length, the average staying time and the average waiting time, etc., especially for the two indexes of the average waiting queue length and the average waiting time. The future research will involve in the area of analyzing this kind of problems from the angle of technical economy, to be specific, the study of the server number to be determined, operational cost and the profit of the queuing system and so on.

Acknowledgement

In this paper, the research was sponsored by the Science and Technology Research and Development Program of Shaanxi Province (project No. 2014JM2-6118).

References

[1] LU Chuanlai. Queuing Theory (Second Edition) [M] Beijing: Beijing University of Posts and Telecommunications Press, 2009

[2] HE Huaiwen, FU Yu, YANG Yihong, XIAO Tao. Service Performance Analysis of Cloud Computing Center Based on Queuing M/M/n/n+r Model [J] Computer Application, 2014, 34 (7): 1843-1847

[3] Wu, K. Taxonomy of batch queuing models in manufacturing systems [J]. European Journal of Operational Research, 2014, 237(1): 129-135.

[4] YAN Yuqing, LI Shixian, SUN Weijun, HUANG Changqin. Study of Requirements Evolution Based on Queuing Theory [J]. Computer Science, 2012, 39 (5）:106-110.

[5] ZHANG Hongbo. The M/M/1 Multiple Queue Model with Bernoulli-controlled Policy [J].

Operations Research Transactions, 2013, 17 (3): 93-100.

[6] ZHANG Hongbo, FENG Pinghua. The M/M/1 Vacation Queue Model with Threshold-controlled Policy [J]. Chinese Journal of Engineering Mathematics, 2013, 30 (4): 569-579.

[7] Kumar, BK; Anbarasu, S; Lakshmi, SRA. Performance Analysis for Queuing Systems with Closedown Periods and Server under Maintenance [J], International Journal of Systems Science, 2015, 46(1): 88-110.

[8] ZHU Yijun, BAO Yuanyuan. M/M/N Repairable Queue System with Non-dominating Priority [J] Systems Engineering and Electronics, 2009, 31(6): 1501-1505.

[9] Ozkan, E; Kharoufeh, JP. Optimal Control of a Two-server Queuing System with Failures [J]. Probability in the Engineering and Informational Sciences, 2014, 28(4): 489-527.

[10] Tao Yang, Wuyi Yue, Jianfen Zhan, et al. Optimization of a Generalized M(k)/M/k Queuing system [J]. International Journal of Revenue Management, 2009, 3(4): 428-440.