日本オペレーションズ・リサーチ学会
2004年秋季研究発表会

Optimal Number of Servers in an M(k)/M/k Queueing System

西安電子科技大学 楊 涛 YANG Tao
上海大学 *胡 奇英 HU Qiying
01107734 甲南大学 岳 五一 YUE Wuyi

1 Introduction

Over the last 30 years, many papers studied different versions of the service mechanisms and the control of the service processes in queueing systems, dealing with both characterization and computation of optimal policies.

Szarkowicz and Knowles [1] discussed the optimal control of an M/M/S queueing system with a controllable number of servers. George and Harrison [2] discussed the dynamic control of an M/M/1 queueing system with adjustable server rate. Andradóttir et al. [3] studied dynamical assignment of the servers to stations in order to obtain optimal long-running average throughput.

Most of the above research assumed that the arrival rate is independent of the number of servers. In this paper, we are concerned with a generalized M(k)/M/k queueing system in which the arrival rate is an increasing function of the number of servers.

2 System Model

We consider an M(k)/M/k queueing system in which the arrival rate depends on the number of servers. The notations are as follows:

k: the number of servers;
$\lambda(k)$: arrival rate, being a function of k;
μ: mean service rate for each server;
$\rho(k) = \dfrac{\lambda(k)}{k\mu}$: the traffic intensity;
$h(n,k)$: holding cost per time unit when n customers are waiting in the queue with k servers in the queueing system;
$S(k)$: service cost rate, representing the expenses for operating k servers per time unit;
$H(k)$: expected holding cost rate, representing the cost paid for customers' waiting per time unit in the steady-state where there are k servers.

It is assumed that the traffic intensity $\rho(k) < 1$ for all $k \geq 1$ and that both $\lambda(k)$ and $S(k)$ are increasing in k. Then the steady probability of n customers in the system is as follows:

$$\pi_n(k) = \begin{cases} \dfrac{1}{n!}k^n\rho^n(k)\pi_0(k), & 1 \leq n \leq k \\ \\ \dfrac{1}{k!}k^k\rho^n(k)\pi_0(k), & n \geq k \end{cases} \tag{1}$$

where

$$\pi_0(k) = \left[\sum_{i=0}^{k-1}\frac{k^i\rho^i(k)}{i!} + \frac{k^k\rho^k(k)}{k!}\cdot\frac{1}{1-\rho(k)}\right]^{-1}.$$

The average number of customers in the queue is

$$\begin{aligned} Lq(k) &= \sum_{n=0}^{\infty}\pi_n(k)(n-k)^+ \\ &= \frac{k^k\rho^{k+1}(k)}{k![1-\rho(k)]^2}\pi_0(k). \end{aligned} \tag{2}$$

We consider that in the steady-state, the total cost rate is the sum of the holding cost rate and the service cost rate as follows:

$$\begin{aligned} C(k) &= S(k) + H(k) = S(k) + \sum_{n=0}^{\infty}\pi_n(k)h(n,k) \\ &= S(k) + \sum_{n=0}^{k-1}\frac{k^n}{n!}\rho^n(k)\pi_0(k)h(n,k) \\ &\quad + \sum_{n=k}^{\infty}\frac{k^k}{k!}\rho^n(k)\pi_0(k)h(n,k). \end{aligned} \tag{3}$$

In the following, we consider the case of a linear holding cost rate, that is, $h(n,k) = h\cdot(n-k)^+$, where $h > 0$ is the holding cost per customer waiting in the queue per time unit. Then, we have $H(k) = h \cdot Lq(k)$ and so the holding cost rate depends upon the average number $Lq(k)$ of customers in the queue.

3 Optimal Number of Servers

It is assumed in this section that $\rho(k)$ is decreasing in k. Then $Lq(k)$ is rapidly decreasing with k. This tells us that when the traffic intensity decreases, the average number of customers in the queue will decrease rapidly. Hence, the holding cost rate $H(k)$ is rapidly decreasing in k. Let

$$g(k) \;=\; \sum_{i=0}^{k-1} \frac{k!}{i!} \frac{[1-\rho(k)]^2}{k^{k-i}\rho^{k+1-i}(k)} + \frac{1-\rho(k)}{\rho(k)} \quad (4)$$

and $\Delta Lq(k) = Lq(k+1) - Lq(k)$. For $\forall k \geq 1$, we have upper and lower bounds of $\Delta Lq(k)$ given by

$$-\frac{\lambda^2(1)}{\mu^2 - \lambda^2(1)} < -\frac{g(k+1)-g(1)}{g(k)g(k+1)} \leq \Delta Lq(k) < 0. \quad (5)$$

According to the above discussions, the holding cost rate $H(k)$ is rapidly decreasing in k while the service cost rate $S(k)$ increases with k. So we should seek a trade-off between these two costs. The objective of the problem is to minimize the total cost rate in the steady state as follows:

$$\min\{C(k) \mid k = 1, 2, \cdots\}. \quad (6)$$

Denote k^* as the optimal number of servers. Namely, k^* is the optimal solution of the problem given in Eq. (6).

Let $\Delta S(k) = S(k+1) - S(k)$ and $\Delta H(k) = H(k+1) - H(k)$. Then

$$\Delta C(k) := C(k+1) - C(k) = \Delta S(k) + \Delta H(k).$$

We have obviously that 1) if $\Delta C(k) \geq 0$ for all k then $k^* = 1$; 2) if $\Delta C(k) < 0$ for all k then $k^* = +\infty$; and 3) if there is $k_0 \geq 1$ such that $\Delta C(k) \leq 0$ for all $k < k_0$ and $\Delta C(k) \geq 0$ for all $k \geq k_0$, then $k^* = k_0$. Moreover, if $\lim_{k \to +\infty} S(k) = +\infty$ then k^* is finite.

In the following, we consider a special case where the service cost rate $S(k) = s_0 + s \cdot k$ is linear for some constants s_0 and $s > 0$. Then from the above result, the optimal number of servers is finite. Let

$$K = \min\left\{k \mid k \in \boldsymbol{N},\ \frac{g(k)g(k+1)}{g(k+1)-g(1)} > \frac{h}{s}\right\}$$

where $\boldsymbol{N}$ is the set of all positive integers. We will show that K is finite. Then the optimal number of servers is less than K and equals 1 when $s > \dfrac{\lambda(1)^2}{\mu^2 - \lambda(1)^2}$.

This gives an upper limit of the optimal number of servers. Due to the definition, K is decreasing in the coefficient s in $S(k)$. When s is larger, the service cost rate is larger. So the optimal number of servers will decrease with s. In fact, the above result says that the optimal number is 1 when s is significantly large.

4 Numerical Results

Here we give some examples to illustrate the model and the results. Assume that $S(k) = s_0 + s \cdot k$ with s_0, $s > 0$. So $H(k) = h \cdot Lq(k)$. We can see that k^* depends only on $\rho(k)$ and the value of $r = s/h$ but is irrespective of the parameter s_0. So we assume that $s_0 = 0$ hereafter.

Next, we compute the optimal number of servers for different parameters and various forms of $\rho(k)$ and compare these results. We consider an example, where the mean arrival rate of customers is a linear function in k: $\lambda(k) = \lambda \cdot k$ with $\lambda > 0$. Then the traffic intensity is $\rho(k) = \lambda(k)/\mu \cdot k = \lambda/\mu$.

We denote that $\rho_0 = \lambda/\mu$ and select various value of r and ρ_0 to compute the optimal number of servers in the system. The results are listed in Table 1. From Table 1, we see that the optimal number of servers k^* increases with ρ_0 and decrease with r. This is still true for several examples of nonlinear traffic intensity we made.

Table 1. Optimal number of servers.

$r =$	10	1.0	0.5	0.2	0.1	0.05
$\rho_0 = 0.1$	1	1	1	1	1	1
$\rho_0 = 0.7$	1	1	1	2	5	10
$\rho_0 = 0.99$	1	1	1	9	35	126

5 Conclusions

In this paper, we have discussed the optimal design problem in an M(k)/M/k queueing system with the arrival rate depending on the number of servers. By analyzing the cost structure, we get that the holding cost rate is rapidly decreasing in relation to the number of servers. Moreover, we present a method for arriving at the optimal number of servers in a queueing system.

Further studies should include research into dynamic control of such queueing systems where the arrival rate depends on the number of servers.

References

[1] Szarkowicz, D. S., Knowles, T. W., "Optimal control of an M/M/S queueing system," *Oper. Res.*, Vol. 33, No. 3, pp. 644-660, 1985.

[2] George, J. M., Harrison, J. M., "Dynamic control of a queue with adjustable server rate," *Oper. Res.*, Vol. 49, No. 5, pp. 720-731,2001.

[3] Andradóttir, S., Ayhan, H. and Down, D. G., "Server assignment policies for maximizing the steady-state throughput of finite queueing systems," *Mgt. Sci.*, Vol. 47, No. 10, pp. 1421-1439, 2001.