J. Operations Research Soc. of Japan
Vol. 18, No.1 & 2, July 1975

OPTIMAL ALLOCATION OF SERVICE RATES FOR MULTI-SERVER MARKOVIAN QUEUE

AKIHIKO TAHARA AND TOSHIO NISHIDA

Osaka University

(Received October 11, 1972; Revised February 5, 1974)

Abstract

Multi-server Markovian queue with no waiting room is considered. An incoming unit is assigned to the server of largest service rate among free ones. And the sum of service rates of all servers is assumed to be constant. Under these conditions, it is shown that the optimal service rates of each server are positive and different from each other. Here 'optimal' is used in the sense of minimizing the rate of loss calls.

A table of optimal allocation of service rates for three server case is attached.

1. Introduction

One of the modern tendencies of queuing theory is to find an optimal design of service system. This paper deals with a multi-server Markovian queuing system with no waiting room and attempts to improve its efficiency. The low efficiency of a queuing system can of course be remedied by rapid service at a large number of servers. This does not require a discussion. Hence it is assumed that the number of servers $c \geq 2$, the arrival rate $\lambda > 0$, and the total service rate $\mu > 0$ are given. The service rate of each server, on the other hand, is to be determined, subject to the constraint that the sum of service rates of all servers is equal to μ, so as to minimize the rate of loss calls (the probability that a unit arrives to find all servers busy and is lost), or equivalently, to maximize the expected number of services per unit time. In other words, the problem is to find the most efficient system among all ones with common c, λ, and μ.

In tackling the above problem, the assignment rule of incoming units comes into question, since a system considered here has in general heterogeneous servers. Of course, if a unit arrives to find only one server free, he is assigned to this free one, and if no servers free, he goes away without being served. When more than one server is free, an incoming unit, according to the ordinary queue discipline, is assigned to each free server at random. But, it will be more efficient to assume an incoming unit to be assigned to the server (or a server) of largest service rate among free ones, since in this case we are using the faster server as often as possible. This assignment rule will be called FSR (fastest server rule). In the case of human customers who know about the differences in the service abilities, FSR will be carried out on their own initiative.

In this paper we shall show that under FSR the optimal values of service rates of c servers are not identical but different from each other. From this, it will be seen that an optimal system obtained here, which has heterogeneous servers and uses FSR, is more efficient than the ordinary M/M/c(c) with homogeneous servers, each with service rate μ/c, and the ordinary queue discipline.

2. Preliminaries

Assume servers are numbered and let μ_i be the service rate of the i-th server, i=1,2,$\cdots$,c. By the assumption,

$$\mu_i \geqq 0 , \quad \mu_1+\mu_2+\cdots+\mu_c = \mu > 0 . \tag{1}$$

Denote by $S=(\mu_1,\mu_2,\cdots,\mu_c)$ a queuing system in which the service rate of the i-th server is given by μ_i. For any S, we assume throughout this paper that arriving units are assigned to the free server of lowest number. But, this will not prevent us from finding the optimal service rates under FSR, since we can suppose without loss of generality that

$$\mu_1 \geqq \mu_2 \geqq \cdots \geqq \mu_c . \tag{2}$$

Let F(S) denote the rate of loss calls of a system $S=(\mu_1,\mu_2,\cdots \mu_c)$. By J.Riordan [1], it follows that

$$F(S) = [\ -\lambda\phi'_{c+1}(0)\ ]^{-1} = \prod_{i=1}^{c} \phi_i(\mu_i)\ , \tag{3}$$

where $\phi_1(s)$ is the Laplace-Stieltjes transform of the arrival distribution;

$$\phi_1(s) = \lambda/(s+\lambda) \tag{4}$$

and

$$\phi_{i+1}(s) = \phi_i(s+\mu_i)/[\ 1-\phi_i(s)+\phi_i(s+\mu_i)\ ]\ ,\quad i=1,2,\cdots,c \tag{5}$$

which represents the Laplace-Stieltjes transform of the overflow distribution from the first i servers.

The following lemma is obvious, but since it will play an important role in our arguments, we state it here.

Lemma 1. $\phi_i(s)>0$, $\phi_i'(s)<0$, and $\phi_i''(s)>0$ for $s\geq 0$, and $\phi_i(0)=1$; $i=1,2,\cdots,c$.

As an extreme case, if $S=(\mu/c,\mu/c,\cdots,\mu/c)$, then F(S) coincides with the expression known as Erlang's loss formula for the ordinary M/M/c(c) with service rate μ/c for each server [1]. This is a matter of course, since assignment rules have no effect on the capacity of queuing system with homogeneous servers. If, on the other hand, $S=(\mu,0,0,\cdots,0)$, then $F(S)=\phi_1(\mu)$, which is the rate of loss calls of M/M/1(1) with service rate μ.

3. The Result

Let $S^*=(\mu_1^*,\mu_2^*,\cdots,\mu_c^*)$ be an arbitrary system that minimizes F(S) defined by (3) subject to (1); such a system does exist, because the class of S's is compact and F(S) is a continuous function of S. It will be shown in the next section that

$$\mu_1^* > \mu_2^* > \cdots > \mu_c^* > 0. \tag{6}$$

Since (6) satisfies (2), S^* is an optimal system; S^* minimizes F(S) subject to (1) and (2). Therefore, when FSR is being used, the optimal service rates of c servers are pisitive and different from each other. Consequently, a system with homogeneous servers or with servers having no service ability is never optimal. The assumption of homogeneous servers may be partly due to the expectation that the homogeneous system will be better than any heterogeneous one with same total service rate, which is not true as stated above.

An optimal system S^* will be found out by applying Lagrange technique;

$$\frac{\partial F(S)}{\partial \mu_1} = \frac{\partial F(S)}{\partial \mu_2} = \cdots = \frac{\partial F(S)}{\partial \mu_c}, \quad \sum_{i=1}^{c} \mu_i = \mu . \tag{7}$$

For two-server case, upon using the above relation we can easily get

$$\mu_1^* = \frac{\sqrt{1+(\mu/\lambda)}}{1+\sqrt{1+(\mu/\lambda)}}\,\mu , \quad \mu_2^* = \frac{1}{1+\sqrt{1+(\mu/\lambda)}}\,\mu . \tag{8}$$

This result was first derived by V.P.Singh [2]. As a numerical example, when $\lambda/\mu=0.5$, we have $S^*=(0.634\mu, 0.366\mu)$ and $F(S^*)=0.1944$. On the other hand, $F(S')=1/5$ and $F(S'')=1/3$, where $S'=(\mu/2,\mu/2)$ and $S''=(\mu,0)$. The following table lists the optimal system $S^*=(\mu_1^*,\mu_2^*,\mu_3^*)$ for three-server case, where S' and S" in the last two columns mean respectively $(\mu/3,\mu/3,\mu/3)$ and $(\mu,0,0)$.

λ/μ	μ_1^*/μ	μ_2^*/μ	μ_3^*/μ	$F(S^*)$	F(S')	F(S")
1.0	0.4260	0.3265	0.2474	0.3424	0.3462	0.5000
0.8	0.4438	0.3240	0.2322	0.2628	0.2684	0.4444
0.6	0.4700	0.3196	0.2105	0.1724	0.1803	0.3750
0.4	0.5120	0.3109	0.1771	0.0805	0.0898	0.2857
0.2	0.5921	0.2886	0.1193	0.0145	0.0198	0.1667

4. Proof of (6)

We now prove that the condition (6) must hold for $S^*=(\mu_1^*,\mu_2^*,\cdots,\mu_c^*)$ that

minimizes F(S) subject to (1). For this purpose, we need only show that for i=1,2,...,c-1,

$$\text{if} \quad \mu_i^*+\mu_{i+1}^* > 0 \quad \text{then} \quad \mu_i^* > \mu_{i+1}^* > 0, \tag{9}$$

because by virtue of (1), $\mu_i^*+\mu_{i+1}^*>0$ for some i, then by (9), $\mu_i^*>\mu_{i+1}^*>0$, from which $\mu_{i-1}^*+\mu_i^*>0$ and $\mu_{i+1}^*+\mu_{i+2}^*>0$ follow, hence again by (9), $\mu_{i-1}^*>\mu_i^*>0$ and $\mu_{i+1}^*>\mu_{i+2}^*>0$, thus $\mu_{i-1}^*>\mu_i^*>\mu_{i+1}^*>\mu_{i+2}^*>0$, and so on.

Assuming $\mu_i^*+\mu_{i+1}^*=2a>0$ for a fixed i, define

$$S(x) = (\mu_1^*,\cdots,\mu_{i-1}^*,x,2a-x,\mu_{i+2}^*,\cdots,\mu_c^*) , \quad 0 \leq x \leq 2a . \tag{10}$$

For each x, S(x) will represent a system in which the service rates of the i-th and the i+1-st servers are respectively given by x and 2a-x. Clearly, S(x) satisfies (1), and $S(\mu_i^*)=S^*$. As to S(x), the notation $\phi_j(s,x)$ will be used for $\phi_j(s)$ defined by (5) in order to clarify its dependence on x, if $j \geq i+1$. Needless to say, if $j \leq i$, $\phi_j(s)$ does not depend on x. Then the expression (3) can be written as

$$F(x) \equiv F(S(x)) = RG(x)H(x) , \quad 0 \leq x \leq 2a , \tag{11}$$

where we put

$$R = \prod_{j=1}^{i-1} \phi_j(\mu_j^*) , \tag{12}$$

$$G(x) = \phi_i(x)\phi_{i+1}(2a-x,x) , \tag{13}$$

and

$$H(x) = \prod_{j=i+2}^{c} \phi_j(\mu_j^*,x) . \tag{14}$$

Since $\mu_i^*+\mu_{i+1}^*=2a>0$, (9) is equivalent to $a<\mu_i^*<2a$. It is therefore sufficient to establish that

$$F'(x) < 0 \quad \text{for} \quad 0 \leq x \leq a \quad \text{and} \quad F'(2a) > 0 . \tag{15}$$

This will immediately follow if we can show

$$G'(x) < 0 \quad \text{for} \quad 0 \leq x \leq a \quad \text{and} \quad G'(2a) > 0 \tag{16}$$

and

$$H'(x) \leq 0 \quad \text{for} \quad 0 \leq x \leq a \quad \text{and} \quad H'(2a) \geq 0 \ , \tag{17}$$

since R, G(x), and H(x) are positive, by virtue of Lemma 1. We first prove (16) and then (17).

Substituting (5) into (13) yields

$$G(x) = \phi_i(x)\phi_i(2a)/[\ 1-\phi_i(2a-x)+\phi_i(2a)\ ] \ . \tag{18}$$

We shall in what follows write $\phi(\cdot)$ for $\phi_i(\cdot)$. Then we have

$$G'(x) = \phi(2a)\psi(x)/[\ 1-\phi(2a-x)+\phi(2a)\ ]^2 \ , \tag{19}$$

where

$$\psi(x) = [\ 1-\phi(2a-x)+\phi(2a)\ ]\phi'(x) - \phi(x)\phi'(2a-x) \ . \tag{20}$$

By Lemma 1, it is easily seen that

$$\psi(a) = [1-2\phi(a)+\phi(2a)\ ]\phi'(a) < 0 \ , \tag{21}$$

$$\psi(2a) = \phi(2a)[\ \phi'(2a)-\phi'(0)\ ] > 0 \ , \tag{22}$$

and for $0 \leq x \leq 2a$,

$$\psi'(x) = [\ 1-\phi(2a-x)+\phi(2a)\ ]\phi''(x) + \phi(x)\phi''(2a-x) > 0. \tag{23}$$

Since $\phi(2a)>0$, we obtain (16).

To prove (17), let $\phi_{j,x}(s,x)$ denote the partial derivative of $\phi_j(s,x)$ with respect to x, for $j \geq i+2$. Since $\phi_j(\mu_j^*,x)>0$, it suffices to see that

$$\phi_{j,x}(\mu_j^*,x) \leq 0 \quad \text{for} \quad 0 \leq x \leq a \quad \text{and} \quad \phi_{j,x}(\mu_j^*,2a) \geq 0 \ ; \ j \geq i+2 \ . \tag{24}$$

But, from (5), we have for $j \geq i+2$,

$$\phi_{j+1,x}(s,x) = \frac{[\ 1-\phi_j(s,x)\ ]\phi_{j,x}(s+\mu_j^*,x) + \phi_j(s+\mu_j^*,x)\phi_{j,x}(s,x)}{[\ 1-\phi_j(s,x)+\phi_j(s+\mu_j^*,x)\ ]^2} \ , \tag{25}$$

where $0<\phi_j(s,x)\leq 1$ for $s \geq 0$, by Lemma 1. This will imply that if the sign of $\phi_{j,x}(s,x)$ at some x remains unchanged over all $s \geq 0$, so does $\phi_{j+1,x}(s,x)$, with the same sign as $\phi_{j,x}(s,x)$. Therefore, in order to prove (24), we need only show that for all $s \geq 0$,

$$\phi_{i+2,x}(s,x) \leq 0 \quad \text{for} \quad 0 \leq x \leq a \quad \text{and} \quad \phi_{i+2,x}(s,2a) \geq 0 \ . \tag{26}$$

Now, by use of (5), it easily follows that for $s \geq 0$ and $0 \leq x \leq 2a$,

$$\phi_{i+2}(s,x) = \frac{\phi(s+2a)[\ 1-\phi(s)+\phi(s+x)\ ]}{[\ 1-\phi(s)\ ][\ 1-\phi(s+2a-x)+2\phi(s+2a)\ ] + \phi(s+x)\phi(s+2a)} \ , \tag{27}$$

where $\phi(\cdot)=\phi_i(\cdot)$ as before. Hence,

$$\phi_{i+2,x}(s,x) = \frac{\xi_3(1-\xi_0)[\ (1-\xi_2+\xi_3)\eta_1-(1-\xi_0+\xi_1)\eta_2\ ]}{[\ (1-\xi_0)(1-\xi_2+2\xi_3)+\xi_1\xi_3\ ]^2} \ , \tag{28}$$

where for notational convenience we put $\xi_0=\phi(s)$, $\xi_1=\phi(s+x)$, $\xi_2=\phi(s+2a-x)$, $\xi_3=\phi(s+2a)$, $\eta_1=\phi'(s+x)$, and $\eta_2=\phi'(s+2a-x)$. Setting

$$\zeta(s,x) = (1-\xi_2+\xi_3)\eta_1 - (1-\xi_0+\xi_1)\eta_2 \tag{29}$$

and noting Lemma 1 we find that

$$\zeta(s,a) = (\xi_0-2\xi_1+\xi_3)\eta_1 < 0 \ , \tag{30}$$

$$\zeta(s,2a) = (1-\xi_0+\xi_3)(\eta_1-\eta_2) > 0 \ , \tag{31}$$

and for $0 \leq x \leq 2a$,

$$\partial\zeta(s,x)/\partial x = (1-\xi_2+\xi_3)\phi''(s+x) + (1-\xi_0+\xi_1)\phi''(s+2a-x) > 0 \ . \tag{32}$$

Applying these relations to (28) gives (26), since $\xi_3(1-\xi_0) \geq 0$ for $s \geq 0$. Consequently, we get (17).

Thus the proof of (6) is complete.

References

[1] J.Riordan, *Stochastic Service System*, J.Wiley, New York, pp.36-40, 1962.

[2] V.P.Singh, "Two-Server Markovian Queues with Balking: Heterogeneous vs. Homogeneous Servers," *Opns. Res.*, 18, 145-159, (1970).