Journal of Engineering Mathematics, Vol. 7, No. 2, April 1973
Noordhoff International Publishing – Leyden
Printed in The Netherlands

Queue-dependent servers

V. P. SINGH

IBM Components Division, East Fishkill, New York 12533, USA

(Received April 4, 1972)

SUMMARY

A Markovian queue with number of servers depending upon queue length is discussed. Whenever the queue in front of the first server reaches a certain length, the system starts another server. There are costs associated with the opening of a new server and the waiting of the customers. A relationship among the costs, traffic intensity and the queue size is obtained.

1. Introduction

In many situations when there are too many people waiting to be served in front of a service facility, the system opens another service facility to reduce congestion. For example, this happens in the banks and at the checkout counters in the department and grocery stores over the weekends.

In this paper we study a Markovian queue in such a way that a new service facility is provided by the system whenever the queue in front of a server reaches a certain length. The different servers may either have the same or different service rates. A new service facility is started at some cost to the system. There is also a cost associated with the difference in the average number of customers in a single server system and the new system. For the case of two homogeneous servers, a relationship is developed among the costs, the traffic intensity ρ and the maximum allowable queue size N in front of the first server. For different values of N and ρ, the ratio of the costs is given in a table. This situation is then discussed for the case of three homogeneous servers. Finally the case of two heterogeneous servers is discussed.

2. The queue $M/M/2$ with number of servers depending on queue length

Customers arrive at a single service counter following an orderly, stationary Poisson stream without after effects with parameter λ. Whenever there are N customers in the queue, the service system starts another server to reduce congestion. It costs c_2 dollars to the system to provide the second server. The service time distribution for each server is negative exponential with parameter μ.

Let P_n denote the steady-state probability that there are n customers in the system at any time. Then the balance equations for the above system take the following form:

$$\lambda P_0 = \mu P_1 ,$$

$$(\lambda+\mu)P_n = \lambda P_{n-1}+\mu P_{n+1} , \quad 1 \leqq n < N ,$$

$$(\lambda+\mu)P_N = \lambda P_{N-1}+2\mu P_{N+1} , \quad n = N ,$$

$$(\lambda+2\mu)P_n = \lambda P_{n-1}+2\mu P_{n+1} , \quad n > N .$$

The solution of the above system of equations is

$$P_0 = \frac{(1-\rho)(2-\rho)}{2-\rho-\rho^{N+1}}, \text{ where } \rho = \frac{\lambda}{\mu},$$

$$P_n = \begin{cases} \rho^n P_0 & 0 \leqq n \leqq N \\ \dfrac{\rho^n P_0}{2^{n-N}} & n > N \,. \end{cases}$$

Prob {second server is operating} = Prob $\{n > N\}$

$$= \sum_{n=N+1}^{\infty} P_n = \frac{\rho}{2-\rho} P_N = \frac{\rho^{N+1}}{2-\rho} P_0 = \frac{(1-\rho)\rho^{N+1}}{2-\rho-\rho^{N+1}} .$$

The average number of customers in the system is given by:

$$E(Q_2) = \sum_{n=0}^{\infty} nP_n = P_0 \sum_{n=0}^{N} n\rho^n + P_0 2^N \sum_{n=N+1}^{\infty} n \left(\frac{\rho}{2}\right)^n$$

$$= \frac{\rho(2-\rho)}{(1-\rho)(2-\rho-\rho^{N+1})} - \frac{(2-\rho)(N-\rho N+1)\rho^{N+1}}{(1-\rho)(2-\rho-\rho^{N+1})} + \frac{(1-\rho)(2N-\rho N+2)\rho^{N+1}}{(2-\rho)(2-\rho-\rho^{N+1})} .$$

Recall that the corresponding expression in a single server system is:

$$E(Q_1) = \rho/(1-\rho) .$$

Let c_1 be the unit cost associated with $E(Q_1)-E(Q_2)$. From the point of view of this cost and the cost in starting a second server, it is profitable for the system to have the second server only if

$$c_1(E(Q_1)-E(Q_2)) > c_2 \text{ Prob } \{n > N\} ,$$

i.e.,

$$c_1\left(\frac{\rho}{1-\rho} - \frac{\rho(2-\rho)}{(1-\rho)(2-\rho-\rho^{N+1})}\right)$$

$$+ c_1 \frac{\rho^{N+1}}{2-\rho-\rho^{N+1}}\left(\frac{(2-\rho)(N-\rho N+1)}{(1-\rho)} - \frac{(1-\rho)(2N-\rho N+2)}{(2-\rho)}\right) > c_2 \frac{(1-\rho)\rho^{N+1}}{(2-\rho-\rho^{N+1})},$$

TABLE 1

Upper bounds for c_2/c_1.

N \ ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	2.28	2.64	3.11	3.75	4.67	6.07	8.46	13.33	28.18
2	3.39	3.89	4.54	5.41	6.67	8.57	11.80	18.33	38.18
3	4.50	5.14	5.96	7.08	8.67	11.07	15.13	23.33	48.18
4	5.61	6.39	7.39	8.75	10.67	13.57	18.46	28.33	58.18
5	6.72	7.64	8.82	10.41	12.67	16.07	21.08	33.33	68.18
6	7.83	8.89	10.25	12.08	14.67	18.57	25.13	38.33	78.18
7	8.94	10.14	11.68	13.75	16.67	21.07	28.46	43.33	88.18
8	10.05	11.39	13.10	15.41	18.67	23.57	31.79	48.33	98.18
9	11.16	12.64	14.53	17.08	20.67	26.07	35.13	53.33	108.18
10	12.28	13.89	15.96	18.75	22.67	28.57	38.46	58.33	118.18
11	13.39	15.14	17.39	20.41	24.67	31.07	41.79	63.33	128.18
12	14.50	16.39	18.82	22.08	26.67	33.57	45.13	68.33	138.18
13	15.61	17.64	20.24	23.75	28.67	36.07	48.46	73.33	148.18
14	16.72	18.89	21.67	25.41	30.67	38.57	51.79	78.33	158.18
15	17.84	20.14	23.10	27.08	32.67	31.07	55.13	83.33	168.18
16	18.95	21.39	24.53	28.75	34.67	43.57	58.46	88.33	178.18
17	20.06	22.64	25.96	30.41	36.67	46.07	61.79	93.33	188.18
18	21.17	23.89	27.38	32.08	38.67	48.57	65.13	98.33	198.18
19	22.28	25.14	28.81	33.75	40.67	51.07	68.46	103.33	208.18
20	23.39	26.39	30.24	35.41	42.67	53.57	71.79	108.33	218.18

$$c_1\frac{(-\rho)}{(1-\rho)}+c_1\left(N(2-\rho)+\left(\frac{2-\rho}{1-\rho}\right)-N(1-\rho)-2\left(\frac{1-\rho}{2-\rho}\right)\right)>(1-\rho)c_2\,,$$

$$N>(1-\rho)\frac{c_2}{c_1}-\frac{2}{2-\rho}\,.$$

Let $c^*=c_2/c_1$, then the above inequality can be written as

$$c^*<\frac{N}{1-\rho}+\frac{2}{(1-\rho)(2-\rho)}\,.$$

It is to be noted that the above relation depends only on c^*, N and ρ. Thus, given any two of these, the best value of the third quantity can be computed. Table 1 gives upper bounds for c^* for specific values of N and ρ.

3. The queue *M/M/3*

We now consider the case in which the system already has two servers and starts a third server. Let N and M are the numbers in the queue when servers 2 and 3 are started up respectively. The balance equations for this system take the following form:

$$\begin{aligned}
&\lambda P_0=\mu P_1\,,\\
&(\lambda+\mu)P_n &&=\lambda P_{n-1}+\mu P_{n+1}\,, && 1\leqq n<N\,,\\
&(\lambda+\mu)P_N &&=\lambda P_{N-1}+2\mu P_{N+1}\,,\\
&(\lambda+2\mu)P_n &&=\lambda P_{n-1}+2\mu P_{n+1}\,, && N<n<M\,,\\
&(\lambda+2\mu)P_M &&=\lambda P_{M-1}+3\mu P_{M+1}\,,\\
&(\lambda+3\mu)P_n &&=\lambda P_{n-1}+3\mu P_{n+1}\,, && n>M\,.
\end{aligned}$$

The following solution to the above system of equations can be verified:

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

$$P_n=\begin{cases}\rho^nP_0 & 0\leqq n\leqq N\\ 2^{N-n}\rho^nP_0 & N<n\leqq M\\ 2^{N-M}3^{M-n}\rho^nP_0 & n>M\end{cases}$$

The average number of customers in the system is:

$$\begin{aligned}
E(Q_3)&=\sum_{n=0}^{\infty}nP_n\\
&=P_0\left[\sum_{n=1}^{N}n\rho^n+2^N\sum_{n=N+1}^{M}n\left(\frac{\rho}{2}\right)^n+2^{N-M}3^M\sum_{n=M+1}^{\infty}n\left(\frac{\rho}{3}\right)^n\right]\\
&=\left[\frac{\rho}{(1-\rho)^2}(1-\rho^N(N-\rho N+1))+\frac{\rho}{(2-\rho)^2}((2N-\rho N+2)\rho^N-(2M-\rho M+2)2^N(\rho/2)^M)\right.\\
&\qquad\left.+\left(\frac{1}{2}\right)^{M-N}\rho^{M+1}\frac{3M-\rho M+3}{(3-\rho)^2}\right]P_0\,.
\end{aligned}$$

Let it cost c_3 dollars to the system to provide the third server and let c_1 be the unit cost associated with the differences in the average number of customers in the single server and two server,

and two server and three server systems. Then it will be profitable for the system to have the third server only if

$$c_1(E(Q_2))-E(Q_3)) > c_2 \text{ Prob}(n > M),$$
$$c_1(E(Q_1)-E(Q_3)) > c_2 \text{ Prob}(N < n \leqq M)+c_3 \text{ Prob}(n > M).$$

4. The queue $M/M_i/2$

The analogous results for two (2) and three (3) heterogeneous servers can be easily obtained. For example, in a two (2) server heterogeneous system, $M/M_i/2$, where μ_1 and μ_2 are the service rates for the first and second server respectively. The balance equations are:

$$\lambda P_0 = \mu_1 P_1,$$
$$(\lambda+\mu_1)P_n = \lambda P_{n-1}+\mu_1 P_{n+1}, \qquad 1 \leqq n < N,$$
$$(\lambda+\mu_1)P_N = \lambda P_{N-1}+\mu P_{N+1},$$
$$(\lambda+\mu)P_n = \lambda P_{n-1}+\mu P_{n+1}, \qquad n > N,$$

and

$$P_n = \begin{cases} \rho_1^n P_0, & 0 \leqq n < N \\ \rho^n(\rho_1/\rho)^N P_0, & n \geqq N \end{cases}$$

where

$$\rho_1 = \lambda/\mu_1, \quad \rho = \lambda/\mu, \quad \mu = \mu_1+\mu_2,$$
$$P_0 = (1-\rho_1)(1-\rho)/(1-\rho-(\rho_1-\rho)\rho_1^N).$$

$$\text{Probability } (n > N) = \frac{\rho}{1-\rho}\rho_1^N P_0,$$

$$E(Q_2) = \left[\rho_1 \frac{1-(N+1-N\rho_1)\rho_1^N}{(1-\rho_1)^2} + (\rho_1/\rho)^N \frac{(N+1-N\rho)\rho^{N+1}}{(1-\rho)^2}\right] P_0.$$

In this case, it is profitable for the system to have the second server operating only when the following inequality is satisfied.

$$N > c^* \frac{\rho(1-\rho_1)}{\rho_1-\rho} \frac{1}{1-\rho}.$$

Acknowledgement

The author wishes to thank Dr. U. N. Bhat for various helpful comments and suggestions during the preparations of this work.