MTA Számitástechnikai és Automatizálási Kutató Intézete, **Közlemények** 17/1976.

OPTIMIZATION PROBLEMS IN A SIMPLE MARKOV SERVICE SYSTEM

by Ahmed F. Mashhour

ABSTRACT

A service system $M(M)1, (n + 1)$ which operates for a finite period of time, is considered. The system is associated with a simple cost structure. The paper deals first with the problem of finding the optimal service rate which minimizes the expected total costs. Then the optimal arrival rate under the same criterion is investigated. Numerical results for both cases are given. The optimal service rate for queuenig systems with infinite operation time is discussed.

INTRODUCTION

This paper is motivated by [1] in which the input of a similar service system is controlled by using a rejection time policy. The present paper deals with the problem of controlling the system by two different approaches. First controlling the system through the service facility by choosing the optimal service rate. Then the system is controlled through the input by choosing the arrival rate optimally.

Consider an $M/M/1, (n + 1)$ queueing system which operates for a finite period of time $(0,T)$. The system starts at time $t = 0$ with no customers. Customers arrive according to a Poisson stream with mean arrival rate λ. An arriving customer enters the system only when the number of the present customers at his arrival is less than $n + 1$. The service times are independent exponentially distributed random variables with mean $1/\mu$. After the closing time T, no new arrivals are accepted and the present customers in the system, if there are any, are to be served in an overtime. The system is associated with the following costs:

i- The cost (loss) per unit time when the server is idle during the period $(0,T)$, is C_I.

ii- The runing cost per unit time when the server is busy during the period $(0,T)$, iis C_B.

iii- The overtime runing cost per unit time (occurs after the closing time T), is C_0.

To avoide trivial cases, we consider only the cases when

$$C_I > C_0 > C_B, \qquad \text{or} \qquad C_0 > C_1 > C_B.$$

In both cases $C_I > C_B$, since a busy server procedure revenue (the system operates economically), while a free server representes loss for the system.

THE EXPECTED TOTAL IDLE PERIOD DURING $(0, t)$

Let $B_k(t)$ denotes the expected total time the system spends, with no customers during the interval $(0,t)$, given that the system has started at the opening time with k customers,

$k = 0,1,\ldots, n+1.$

If the system starts with k customers at the opening time $t = 0$, then it may happen that the first transition in the Markovian queue size process:

i- is due to an arrival during $(x, x+dx)$ which accurs with probability

$$\lambda e^{-(\lambda+\mu)x}\,dx, \quad \text{if} \quad 0 \leqslant k \leqslant n.$$

ii- is due to a departure during $(x, x+dx)$ which occurs with probability

$\mu e^{-\mu x}\,dx$ if $k = n+1$ and with probability

$$\mu e^{-(\lambda+\mu)x}\,dx \quad \text{if} \quad 1 \leqslant k \leqslant n.$$

Integrating over all values of $0 \leqslant x \leqslant t$, it follows taht $B_k(t)$, $k = 0,1,\ldots, n+1$ satisfies the system of integral equations

$$(1) \quad \begin{cases} B_0(t) = \lambda \int_0^t [x + B_1(t-x)]e^{-\lambda x}\,dx, \\ B_k(t) = \lambda \int_0^t e^{-(\lambda+\mu)x} B_{k+1}(t-x)dx + \mu \int_0^t e^{-(\lambda+\mu)x} B_{k-1}(t-x)dx, \quad 1 \leqslant k \leqslant n, \\ B_{n+1}(t) = \mu \int_0^t e^{-\mu x} B_{n+1}(t-x)dx. \end{cases}$$

If $B_k^*(s) = \int_0^\infty e^{-st} B_k(t)dt$, $k = 0,1,\ldots, n+1$ denote the Laplace transform of $B_k(t)$,

then the system (1) can be written in the form

$$(2) \quad \begin{cases} (\lambda+s)B_0^*(s) - \lambda B_1^*(s) = \lambda/s(\lambda+s), \\ (\mu+s)\,B_{n+1}^*(s) - \mu B_n^*(s) = 0, \\ (\lambda+\mu+s)B_k^*(s) - \lambda B_{k+1}^*(s) - \mu B_{k-1}^*(s) = 0, \quad 1 \leqslant k \leqslant n, \end{cases}$$

The determinánt of the coefficients of the system (2) has the form

$$\Delta_{n+2}(s) = \begin{vmatrix} \lambda+s & -\lambda & 0 & \cdots & & 0 \\ -\mu & \lambda+\mu+s & -\lambda & \cdots & & 0 \\ 0 & & & & & \vdots \\ \vdots & & & & & \vdots \\ \vdots & & & & & 0 \\ & & & -\mu & \lambda+\mu+s & -\lambda \\ 0 & \cdots & & 0 & -\mu & \mu+s \end{vmatrix}$$

Denoting the right lower subdeterminant of order k in $\Delta_{n+2}(s)$ by $\Delta_k(s)$, then it can be easily shown that

(3) $$B_0^*(s) = \frac{\lambda\Delta_{n+1}(s)}{s(\lambda + s)\Delta_{n+2}(s)}$$

In order to decompose $B_0^*(s)$ by partial fractions, we examine the roots its denumerator

$$p_{n+4}(s) = s(\lambda + s)\,\Delta_{n+2}(s).$$

It can be shown, as in [2], that $P_{n+4}(s) = 0$ has

a) one repeated root $s_0 = 0$,
b) one root $s = -\lambda$,
c) $n + 1$ distinicit negative roots $s_1, s_2, \ldots, s_{n+1}$.

It is easy to show that the necessary and sufficient condition that one of the roots $s_1, s_2, \ldots, s_{n+1}$ coincides with the single root $s = -\lambda$ is

$$\Delta_n(-\lambda) = 0.$$

By the virtue of the above discussion, if $\Delta_n(-\lambda) \neq 0$, then

$$p_{n+4}(s) = (s - s_0)^2 (s + \lambda) \prod_{i=1}^{n+1} (s - s_i),$$

and

(4) $$B_0^*(s) = \frac{b_0}{s^2} + \frac{a_0}{s} + \frac{c}{s + \lambda} + \sum_{j=1}^{n+1} \frac{b_j}{s - s_j},$$

where the coefficients are given by

$$
(5)\quad \begin{cases} b_0 = \dfrac{\mu^{n+1}}{\left|\prod_{i=1}^{n+1} s_i\right|} \qquad c = \dfrac{\Delta_{n+1}(-\lambda)}{\lambda \prod_{i=1}^{n+1}(-\lambda - s_i)}, \\[2ex] b_j = \dfrac{\lambda \Delta_{n+1}(s_j)}{s_j^2 (s_j + \lambda) \prod_{\substack{i=1 \\ i \neq j}}^{n+1} (s_i - s_j)}, \qquad j = 1,2,\ldots,n+1, \\[2ex] \text{and} \\ a_0 = -\left(c + \sum_{j=1}^{n+1} b_j\right). \end{cases}
$$

On inversion, we get

$$
(6)\qquad B_0(t) = b_0 t + a_0 + c e^{-\lambda t} + \sum_{j=1}^{n+1} b_j e^{s_j t}.
$$

By the same way, a similar expression can be obtained for $B_0(t)$ when $\Delta_n(-\lambda) = 0$.

Now the expected total idle and busy period during the time of operation $(0,T)$, is $B_0(T)$ and $T - B_0(T)$ respectively.

THE EXPECTED TOTAL COSTS

It remains now to find the expected overtime caused by the customers present at the closing time T. Let $p_k(t)$, $0 \leqslant k \leqslant n+1$ be the probability that there are k customers at time t, in the system. They satisfy a finite system of linear differential equations. The eigenvalues of that system are the roots of $\Delta_{n+2}(s) = 0$, discussed in section 2. The corresponding eigenvectors can be determined, as in Lemma 2 in [2], to get $p_k(t)$ finally in the form

$$
(7)\qquad p_k(t) = \sum_{i=0}^{n+1} d_i \alpha_{k+1}^{(i)} e^{s_i t}, \qquad k = 0,1,\ldots,n+1,
$$

where $\alpha_{k+1}^{(i)}$ is the $(k+1)^{\underline{th}}$ componant of the eigenvector corresponding to the eigevalue s_i, and $d_i s$ are arbitrary constants to be determined from the initial condition of the system (the number of customers in the system at $t = 0$).

Now the objective function, given that the system has started with no customers, is given by

$$
(8)\qquad C_T(\mu) = C_I B_0(T) + C_B(T - B_0(T)) + C_0 \sum_{i=1}^{n+1} \frac{i}{\mu} p_i(T)
$$

The numerical results concerning the optimal service rate μ^* for fixed values of λ in the case of finite waiting room with capacity n, can be summerized as follows:

a.) $C_I = 4$, $C_0 = 2$, $C_B = 1$, $n = 5$ and $T = 10$,

λ	0.5	1.0	1.5	2.0
μ^*	0.84	1.20	1.50	1.90

b.) $C_I = 2$, $C_0 = 3$, $C_B = 1$, $n = 5$ and $T = 10$,

λ	0.5	1.0	1.5	2.0
μ^*	1.70	2.13	2.60	3.0

Concerning the optimal arrival rate λ^* for fixed values of μ in case of finite waiting room with capacity n, we get

c.) $C_I = 2$, $C_0 = 3$, $C_B = 1$, $n = 5$ and $T = 10$,

μ	2.5	3.0	3.5	4.0
λ^*	2.11	3.05	4.16	5.13

d.) $C_I = 4$, $C_0 = 2$, $C_B = 1$, $n = 5$ and $T = 10$,

μ	0.84	1.20	1.50	1.90
λ	1.08	5.60	7.51	10.0

Tables 1 and 4 shows that for a queueing system $M/M/1$, $(n + 1)$ with fixed values of C_I, C_0, C_B, n and T, the optimal arrival rate λ^* corresponding to a fixed value μ, does not imply that μ is the optimal service rate for the same system with $\lambda = \lambda^*$.
This is due to the fact that the dependence of the objective function (8) on λ and μ is not only throuhg the ratio λ/μ.

OPTIMAL SERVICE RATE FOR QUEUEING SYSTEM WITH $T = \infty$

Consider the system $M/M/1$, $(n + 1)$ which operates for infinite period of time ($T = \infty$). The arrival and service rates are λ and μ respectively. The system is associated with the following costs:

i- r_1 is the revenue provided by a served customer.

ii- r_2 is the loss of the system caused by a lost customer (because of the fullness of the waiting room).

iii- C_I is the cost (loss) per unit time when the server is idle.

Denote by $\nu_A(T)$ the number of the admitted (joining) customers during a time interval $(0,T)$, and by $\nu_L(T)$ the number of lost customers (beceause of the fullness of the waiting room) during $(0,T)$.

The purpose of this section is to find the optimal service rate μ^* that maximizies the average expected net revenue given by

$$(9) \qquad C_1(\mu) = \lim_{T\to\infty} \frac{1}{T}[r_1 E\nu_A(T) - C_I B_0(T)],$$

where $B_0(T)$ is the expected total idle period during $(0,T)$ given by equation (6). Putting $\rho = \lambda/\mu$, then the stationary probabilities p_k^* that there are k customers in the system are given, see [3], by

$$(10) \qquad p_k^* = \frac{1-\rho}{1-\rho^{n+2}}\rho^k = \frac{\rho^k}{1+\rho+\ldots+\rho^{n+1}}, \qquad k = 0,1,\ldots,n+1.$$

Now we have that

$$\lim_{T\to\infty} \frac{E\nu_A(T)}{\lambda T} = 1 - p_{n+1}^*,$$

and

$$(11) \qquad \lim_{T\to\infty} \frac{B_0(T)}{T} = p_0.$$

From equations (10) and (11), the objective function given by (9) can be written in the form

$$(12) \qquad C_1(\mu) = \lambda r_1 \frac{1+\rho+\ldots+\rho^n}{1+\rho+\ldots+\rho^{n+1}} - C_I \frac{1}{1+\rho+\ldots\rho^{n+1}} =$$

$$= \lambda r_1 - \frac{\lambda r_1 \rho^{n+1} + C_I}{1+\rho+\ldots+\rho^{n+1}} \quad .$$

Taking the first derivative of (12) with respect to μ and equating to zero, we get

$$(13) \qquad \lambda r_1 [\rho^{n+1} \sum_{k=0}^{n-1} (k+1)\rho^k - (n+1)\rho^n \sum_{k=0}^{n} \rho^k] + C_I \sum_{k=0}^{n} (k+1)\rho^k = 0.$$

The left hand side of the later equation is a polynomil of degree $2n$ in ρ, it can be written in the form

$$(14) \qquad f_{2n}(\rho) = \sum_{i=0}^{2n} a_i \rho^i,$$

where

$$\begin{aligned} a_i &= (i+1)C_I, & 0 \leqslant i \leqslant n-1, \\ &= (n+1)(C_I - \lambda r_1), & i = n, \\ &= -(2n+1-i)\lambda r_I, & n+1 \leqslant i \leqslant 2n. \end{aligned}$$

It is easly seen that the equation $f_{2n}(\rho) = 0$ has only one positive root $\bar{\rho}$ as follows: Since $f_{2n}(0) = a_0 > 0$ and $f_{2n}(\infty) < 0$, then $f_{2n}(\rho) = 0$ has at least one positive root. Applying Descartes'rule of sings, see [4], (which states that number of positive roots of a polynomial is equal to the number of variations in sign in the sequence of coefficients of this polynomial or is less by an even number), it follows that $f_{2n}(\rho) = 0$ has only one positive root $\bar{\rho}$. It is clear that the number of variations in sign of the coefficients a_is given by (14), does not change whatever the relation betwe en λr_1 and C_I.

Taking the second derivative of the objective function given by (12) with respect to μ, we get

$$\frac{d^2}{d\mu^2} C_1(\mu) \Big|_{\rho=\bar{\rho}} < 0.$$

By virtue of the above discussion we conclude that the optimal service rate μ^* at which the objective function $C_1(\mu)$ attains its maximum is unique and equal to $\lambda/\bar{\rho}$ where $\bar{\rho}$ is the unique positive root of (13).
However the upper bound of the positive root of (13) which is given in [4] by

$$(15) \qquad 1 + \sqrt[(n+1)]{nC_I/\lambda r_1}$$

may give a rough description of the behavior of the unique root $\bar{\rho}$ and concequently the optimal value μ^* of the service rate, when the values of λ, r_1 and C_I changes. We discuss in the following example the behavior of the optimal service rate μ^* for the simple case $n = 1$, where an explicit formula for the unique positive root $\bar{\rho}$ exists.

Example: For waiting room capacity $n = 1$, we get

$$C_1(\mu) = \frac{\lambda r_1 (1+\rho) - C_I}{1 + \rho + \rho^2}$$

and equation (13) gives

$$\lambda r_1 \rho^2 + 2(\lambda r_1 - C_I)\rho + C_I = 0.$$

The positive root $\bar{\rho}$ of the later equation is given by

$$(16) \qquad \bar{\rho} = \frac{-(\lambda r_1 - C_I) + \sqrt{(\lambda r_1 - C_I)^2 - \lambda r_I C_I}}{\lambda r_1} \qquad \text{if } \lambda r_1 \neq C_I$$

$$\bar{\rho} = 1 \qquad \text{if } \lambda r_1 = C_I.$$

From (16) it can be aesily shown that μ^* has the properities

a.) $\frac{d}{dC_I} \mu^* \leqslant 0$,

i.e. the optimal mean service time $1/\mu^*$ increases as C_I increases.

b.) $\frac{d}{dr_I} \mu^* > 0$,

i.e. the optimal mean service time $1/\mu^*$ decreases as r_1 increases.

c.) $\frac{d}{d\lambda} \mu^* > 0$,

i.e. the optimal mean service time $1/\mu^*$ decreases as λ increases.

It is clear that the properties of the optimal mean service time $1/\mu^*$ agrees with the properties of the upper bound of $\bar{\rho}$ given by (15) when $n = 1$.

The numerical results obtained for the optimal service rate μ^* in the case of infinite operation time $(T = \infty)$ can be summerized as follows:

1.) For fixed $C_I = 3$, $r_1 = 2$ and $n = 4$

λ	1	2	3	4
μ^*	0.80	2.04	3.51	5.15

2. For fixed $\lambda = 2$ and $n = 5$, the values of μ are given in the following table

r_1 \ C_1	1	2	3	4
1	2.28	1.88	1.68	1.55
2	2.75	2.28	2.04	1.88
3	3.05	2.55	2.28	2.11
4	3.28	2.75	2.47	2.28

It is clear from tables 1 and 2 that the properties a,b and c for the case $n = 1$, are still the same for larger values of the waiting room capacity ($n = 4$ and $n = 5$). In table 2, μ on the diagonal assume a fixed value ($\mu^* = 2.28$) this is due to the fact that, when

$r_1 = C_I$ then the optimal service rate μ^* depends only on λ and n (see equation 13).

Remark: Let us consider the objective function

$$C_2(\mu) = \lim_{T\to\infty} \frac{1}{T} [C_I B_0(T) + r_2 E\nu_L(T)], \tag{17}$$

which represents the average expected cost rate. Now our purpose is to choose the optimal service μ which minimizes $C_2(\mu)$.
It can be seen that

$$C_2(\mu) = \frac{C_I + \lambda r_2 p^{n+1}}{1 + \rho + \ldots + \rho^{n+1}} \tag{17'}$$

since

$$\lim_{T\to\infty} \frac{E\nu_L(T)}{\lambda T} = p^*_{n+1}.$$

Comparing $C_1(\mu)$ and $C_2(\mu)$, (given by equations 12 and 17), we can see that, if $r_2 = r_1$, then the optimal service rate μ^* that maximizes $C_1(\mu)$ as the same that minimizes $C_2(\mu)$. On the other hand if $r_2 \neq r_1$, then the optimal setvice rate μ^* that minimizes $C_2(\mu)$ is unique and has the same properities a,b and c (replacing r_1 by r_2) described in the given example.

ACKNOWLEGEMENT

I would like to express my deep thanks to my supervisor Dr J. Tomkó for his helpful criticism and suggestions.

References

[1] A.Mashhour – J.Tomkó: Controlling the input for Markovian service systems. Reprents of the Stochastic Control Conference, Budapest, (1974).

[2] A.Mashhour: A first passage problem for an M/M/1 queue. MTA SzTAKI, Közlemények 12. (1974).

[3] L.Takács: Introduction to the Theory of Oueues. Oxford University press. New York. (1962).

[4] A.Kurosh: Higher Algebra. MIR Pubilishers. Moscow, (1972). pp. 247.

Összefoglaló

Optimalizálási feladatok Markov tipusú kiszolgálási rendszerekben

Ahmed F. Mashhour

Markov tipusú M/M/1 kiszolgáló rendszer érkezési intenzitásának optimális megválasztását vizsgálja a jelen dolgozat. Feltételezzük, hogy az üzemeltetési költség (időegységenként) C_B ha foglalt a kiszolgáló, C_I ha szabad, s ha véges a müködési idő és túlóra is fellép, akkor a túlórázás időegységenkénti dija C_0. Az alábbi esetek lehetnek érdekesek:

$$C_I > C_0 > C_B \quad \text{és} \quad C_0 > C_I > C_B.$$

Véges és végtelen müködési időre kiszámitjuk a minimális (végtelen idő esetén az egységnyi időre eső) üzemeltetési költséget biztositó érkezés intenzitást. Az eredményeket számitástechnikai szempontból is analizáljuk s szemléltető numerikus eredményeket közlünk.

Резюме

Задачи оптимизации в простейщих системах обслуживания

Ахмед Ф. Машххоур

В работе исследуется оптимальное определение интенсивности входящего потока простейщих системы (M/M/I) обслуживания. Пусть стоимость работы системы /в единицу времени/ C_B если обслуживающих прибор занят , C_I если свободен. Если время функционирования системы конечно и возникает сверхурочная работа, тогда стоимость сверхурочной работы за единицы времени C_O. Интересны следующие случаи:

$$C_I > C_O > C_B \quad \text{и} \quad C_O > C_I > C_B$$

Приведены иллюстративные нумерические экземпляры.