RESEARCH NOTE

OPTIMAL CONTROL OF SERVICE RATE IN A SERVICE CENTER WITH LAPSE

R. Tavakkoli-Moghaddam

Department of Industrial Engineering, Faculty of Engineering
University of Tehran, Tehran, Iran, tavakoli@ut.ac.ir

A. Azaron

Department of Industrial Engineering, Faculty of Engineering
Bu-Ali Sina University, Hamadan, Iran, aazaron@dal.ca

L. Mehrad-Pay

Department of Industrial Engineering, Mazandaran University of Science and Technology
Babol, Iran, Laleh_mp@yahoo.com

(Received: February 5, 2003 - Accepted in Revised Form: June 10, 2004)

Abstract The purpose of this paper is to analyze the effect of a particular control doctrine applied to the service mechanism of a queuing process with lapse. It is assumed that the service discipline is FCFS (first come, first served), arrival process is Poisson, service time distribution is exponential, service process is one phase and the capacity is infinite. It is also assumed that the customer may give up joining the system when the queue is overcrowded. Expressions are obtained for queue length probabilities for describing control performance. The aim of which is to decrease customer's expectancy time via incorporation of a service cost structure. The model is executed by two control methods, namely the single level control and double level hysteretic control. Finally, the results are compared with each other through solving a numerical example.

Key Words Queuing Systems with Lapse, Optimal Control, Single Level, Double Level, Hysteretic Controls

چکیده این مقاله در زمینه کنترل نرخ بهینه سرویس دهی در یک محیط بانکی است. در این محیط فرض می‌شود که نظام سرویس FCFS (First Come, First Served) است و مشتریان طی یک مرحله سرویس دهی می‌شوند. در این حالت ظرفیت بانک نامحدود است و مسئله انصراف به چشم می‌خورد، یعنی وقتی صف طولانی باشد ممکن است مراجعین دلسرد شوند و مایل نباشند منتظر بمانند. در این مقاله، طرح یک سیستم سرویس دهی بهینه ارائه می‌شود و هدف این است که تعدیل زمان انتظار متقاضی بر حسب ساختار ذاتی هزینه ها صورت پذیرد. با بدست آوردن هزینه انتظار می‌توان برای تعیین تعداد بهینه باجه های فعال و تعیین نرخ هایی که این باجه ها باید بر اساس آن راه اندازی شوند استفاده کرد. برای انجام این طرح، از دو مدل تک سطحی و تشنجی دو سطحی استفاده می‌شود و سپس بنحوی که حداقل هزینه بدست آید با یکدیگر مقایسه می‌شوند. به علت بزرگ شدن حجم و زمان محاسبات، از برنامه کامپیوتری که به زبان پاسکال نوشته شده، برای این منظور مورد استفاده قرار می‌گیرد.

1. INTRODUCTION

One of the inevitable facts of life is waiting in queues that may take up an important part of daily routine time. Unfortunately, the importance of this phenomenon is growing with increasing the population, causing waste of time, energy, and money. Although we can never get rid of queues, we can alleviate its advance effects as much as possible. The queuing theory tries to reduce the

waiting time in queue, based on an extensive mathematical analysis.

The queuing theory was first developed to cope with the overcrowding of the telecommunication systems. The first researcher in this area was the Danish mathematician Erlang, who developed a probabilistic theory using in telephone conversations. There are many important applications of this theory, which are appeared in articles related to probability, operations research, and management sciences [1,2].

Since banking systems have been of interest for a long time and banks have always been in competition, special issues such as reduction waiting time and quickness in providing services have been subjects of research. The results of which could he used in determining optimum rate of service-provision, number of the service providers to decrease service times with the least possible amount of costs. By determining the above factors, one can set down a strategic plan for the success of each bank in various dimensions. Along this line, some of the goals would he as follows:

1- Number of the tellers to be employed in this bank.
2- The smallest and the largest number of customers in the queue, on the basic of which the service rate is changed.
3- Minimizing service and waiting costs.

In general, models are divided in two kinds: descriptive and operational [1,2]. Descriptive models are those that explain the present situation, while the operational models provide an optimum behavior of the existing system. Various papers have been presented on this subject.

Hillier and Lieberman [1] believed that if the number of the tellers at a service box is proportionate to service rate, it is better to have one big cluster of tellers (M/M/1), in which the number of tellers is corresponding to μ^* instead of many small clusters of tellers. Of course this is true when the assumption of linearity cost function holds true. One of the works carried out by Stidham [3] proved the queuing formula $L = \lambda \times W$, and its final result which is being used in most optimization models of the queuing systems. Crabill [4] developed various models such as static design

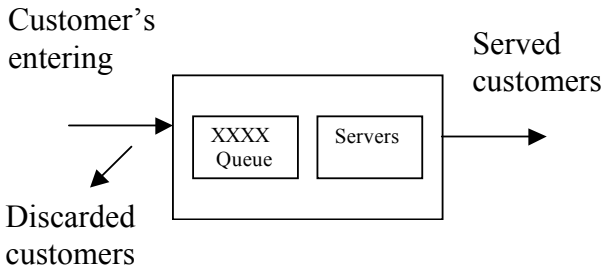

Figure 1. Queuing system with lapse.

models and dynamic control models. Yadin and Naor [5] stated that the service rate can be brought under control of the decision maker at any time.

Gebhardt [6] investigated two special methods to determine the optimum service rate, assuming Poisson arrival and exponential service. Sobel [7] investigated some policies to start and stop the service in a queuing system that affects the service profit. Tembe and Welff [8] investigated the optimum control of service-provision in consequent queues. One of the most recent works on the group entrance (arrival) and exist by group in queuing networks belong to Chao [9]. Crabill [10] presented a paper on the control of service-provision assuming exponential service time and fixed entrance rate. Tijms [11] proposed a policy for a priority-oriented queuing when the service provider can be removed.

2. CREATING A QUEUING SYSTEM

Let a system for servicing exist and customers refer for receiving service. If server is idle, the customer will receive his service immediately, but if server is not idle the customer enters in the queue. The way of creating a queue system has been illustrated in Figure 1. The elements of a queuing system are listed below:

- model of entering customer (A),
- model of servicing (B),
- number of servers (X),
- queue's capacity (Y),
- system array (Z).

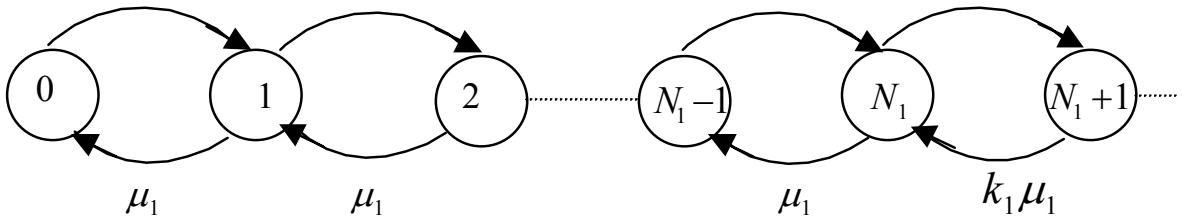

Figure 2. Single level control.

In Markovian queues, the time is divided into three periods of "past", "present" and "future", in which the future is independent of the path it has taken in the "past". It only depends on its situation at the "present" [1, 16]. The size of each queuing system can increase (by birth), or decrease (by death) at any moment. In a queuing system, the customers represent the above population. In this system, the entrance follows a Poisson process with parameter λn, and the service follows another Poisson process with parameter μn. The time interval between two births or two deaths follows an exponential distribution. If we consider the entrance of a particular customer into the system as a birth and his/her exit as a death, then we will have a birth-death process. In this case, the transition rate is shown in Figure 2. The transition rate of the birth-death model shows that each state is related to its previous and latter states and consequently the following relations hold:

$$P_0\ (1+\sum_{i=1}^{\infty} C_i)=1 \Rightarrow P_0=(1+\sum_{i=1}^{\infty} C_i)^{-1}$$
$$P_n=C_nP_0,\quad C_n=\frac{\lambda_0\cdot\lambda_1\cdots\lambda_{n-1}}{\mu_1\mu_2\ldots\mu_n} \tag{1}$$

3. MATHEMATICAL MODEL

Gebhardt [6] investigated a simple M/M/1 queuing system with state-dependent service, in which the arrival process to the system is Poisson and the service time follows an exponential distribution. He considers two models, namely single level control and double level hysteretic control. In the single level control model, the service rate is assumed to be equal μ_1, and it is used while the queue length is less than or equal N_1. Otherwise, the service rate is increased to $k_1 \times \mu_1$, where k_1 is the number of servers.

In the double level hysteretic model, two service rates are considered. As long as the queue length is equal or less than N_1, the service rate $\mu_1$is used. With increasing the queue length from N_1 to N_2, the service rate is also increased from μ1 to $k_1 \times \mu_1$. It means that the number of servers is increased by k_1-1 servers. When the queue length goes back to N_1, then the service rate also returns to μ_1. In fact, the extra servers are removed from the process. One can talk about the existence of a control loop.

Using the above analysis and birth-death process, Gebhardt investigated an efficient method to compute P_0 (the probability that there is no customer in the system) and the other queue length probabilities. The corresponding formulas are shown below. In these formulas λ is the arrival rate, n is the length of queue, k_1 is the number of servers, ρ is the ratio of λ to μ_1 and ρ_1 is the ratio of ρ to k_1. P is also defined as N_2-N_1 = P+1.
In the single control level, we have:

$$P_n=\rho^n P_0 \qquad 0\le n\le N_1-1$$
$$P_n=\rho^{N_1}\rho^{n-N_1}P_0 \qquad 0\le n\le N_1-1 \tag{2}$$

P_0 can be determined from the following

expression:

$$\frac{1}{P_0}=\frac{1}{(1-P)}\left\{1-P^{N_1}\frac{(\rho-\rho_1)}{(1-\rho_1)}\right\} \tag{3}$$

In the double level hysteretic control, we have the following equation.

$$P_n=\rho^n P_0 \qquad 0\le n\le N_1-1$$

$$P'_{N_1+i}=\frac{\rho(\rho^i-\rho^{P+1})P_0}{(1-\rho^{P+1})} \qquad 0\le i\le P$$

$$P'_{N_1+p+j}=\frac{\rho^{N_1+p}\rho_1(1-\rho^{P+1})P_0}{(1-\rho^{P+1})(1-\rho_1)}(I-\rho_1^j)$$
$$1\le j\le P+1$$

$$P_{N_1+p+j}=\frac{\rho^{N_1+p}(1-\rho)(1-\rho_1^{P+1})}{(1-\rho^{P+1})(1-\rho_1)}\rho_1^i P_0$$
$$2\le i\le\infty \tag{4}$$

Primes will be used on those P's that are not the total probabilities of the queue lengths having the subscript values.

Considering the set of Equations 4, it is found the following equation.

$$\frac{1}{P_0}=\frac{1}{(1-\rho)}-\frac{(P+1)\rho^{N_1+P}(\rho-\rho_1)}{(1-\rho^{P+1})(1-\rho_1)} \tag{5}$$

Two appropriate criteria are suggested to measure the performance of double level hysteretic control.

1- The average rate of switching from μ_1 to $k_1\times\mu_1$ denoted by η.
2- The proportion of the total time that the queuing system operates at the greater rate of the two service rates denoted by F.

The switching rate is the rate at which transitions occur from queue length N_1-1 to N_2 or from N_1+1 to N_1. It is found in the following equation.

$$\eta=\lambda P'_{N_1+P}=\lambda\frac{\rho^{N_1+P}(1-\rho)}{(1-\rho^{P+1})}P_0 \tag{6}$$

Equation 6 shows the change of mean rate from μ_1 to $k_1\times\mu_1$. F is equal to the probability of operating at the service rate $k_1\times\mu_1$ and is found as follows:

$$F=\frac{(P+1)\rho^{N_1+P}\rho_1(1-\rho)P_0}{(1-\rho^{P+1})(1-\rho_1)} \tag{7}$$

In this paper, the cost function is comprised of two components as Cs and Cq, which denote the service cost and the queuing cost, respectively. These costs are defined as follows:

$$C_s=c_1\mu+r_1c_1(k_1-1)\mu F+r_2c_1\eta \tag{8}$$

where c_1 is a standard unit of cost. The cost of the additional servers and switching for the additional servers will be varied by r_1 and r_2 in the cost formula, respectively.

With c_2 representing another standard unit of cost, Gebhardt proposed the following forms for Cq:

$$\left\{\begin{array}{ll} Q_1: & C_q=c_2\sum_{n=1}^{\infty}nP_n \\ Q_2: & C_q=c_2\sum_{n=1}^{\infty}(n-5)^2P_n \\ Q_3: & C_q=-3c_2\sum_{n=1}^{\infty}P_n+10c_2\sum_{n=N_2+1}^{\infty}P_n \\ Q_3: & C_q=c_2\left|2.5-E(n)\right|^2+c_2Var(n) \end{array}\right\} \tag{9}$$

Gebhardt goes on to find the cost function for different scenarios. Gebhardt then uses the sum of

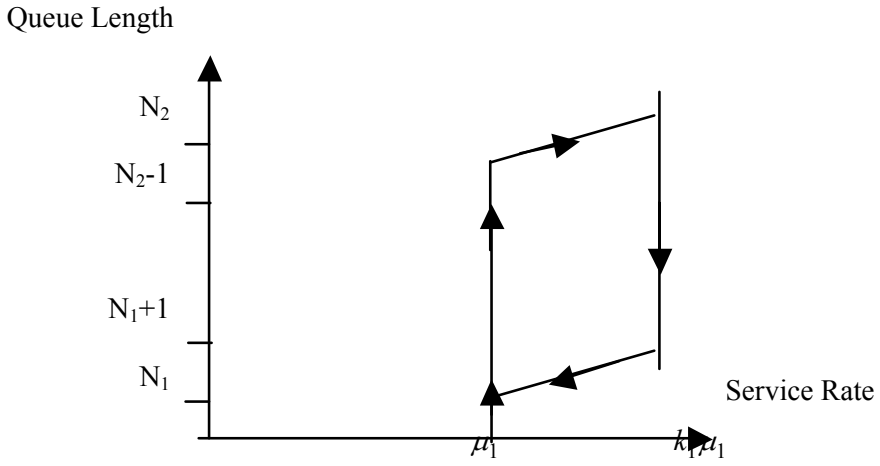

Figure 3. Double level hysteretic control.

Cs and *Cq* to find the strategy of minimum cost.

4. THE PROPOSED MODEL

The model suggested in this paper is a generalized from of the Gebhardt's model [6], in which the lapse theory is also included in the model. It is assumed that the customer may give up entering the queue when the queue seems to be overcrowded. As a result, the arrival rate would be changing and for each state i, we have:

$$\lambda_i = \lambda b_i \tag{10}$$

where b_i is a descending function and is defined as:

$$b_i = 1/(i+1) \tag{11}$$

Considering λ as the arrival rate to the system, μ_1 as the service rate and k_1 as the number of servers in the system, we have:

$$\rho = \frac{\lambda}{\mu_1}, \quad \rho_1 = \frac{\rho}{k_1} \tag{12}$$

It should be noted that this queuing process is always stable, even if the values of ρ or ρ_1 are greater than one, because of considering the lapse in the proposed model.

4.1. Single Level Model In this model, it is supposed that the service rate is equal to μ_1 for the queue lengths less than or equal to N_1 and $k_1 \times \mu_1$ for the queue lengths greater than N_1. Figure 2 shows a single level control. According to Figure 2, we have the following equation.

$$\begin{cases} P_n = \rho^n \prod_{i=1}^{n} b_{i-1} P_0 = \dfrac{\rho^n}{n!} P_0 & 0 \le n \le N_1 - 1 \\ P_n = \rho^{N_1} \rho^{n-N_1} \prod_{i=1}^{n} b_{i-1} P_0 = (\dfrac{\rho}{\rho_1})^{N_1} \dfrac{{\rho_1}^n}{n!} P_0 & \\ N_1 \le n \le \infty & \end{cases} \tag{13}$$

and then we can calculate P_0 as follows:

$$P_0 = \left[\sum_{n=0}^{N_1-1} \frac{\rho^n}{n!} + (\frac{\rho}{\rho_1})^{N_1} \sum_{n=N}^{\infty} \frac{\rho_1^n}{n!} \right]^{-1} \tag{14}$$

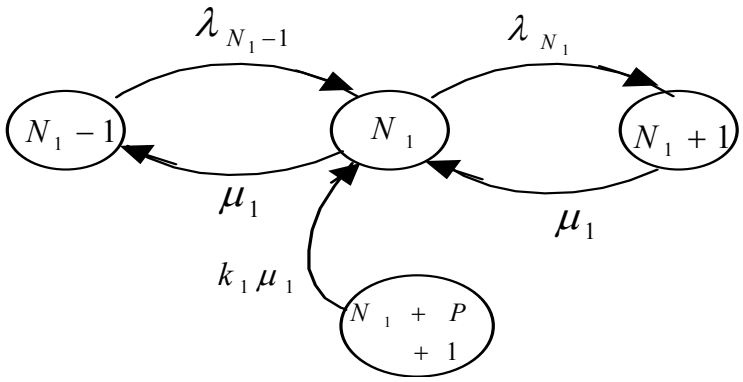

Figure 4. Process in state N_1+1.

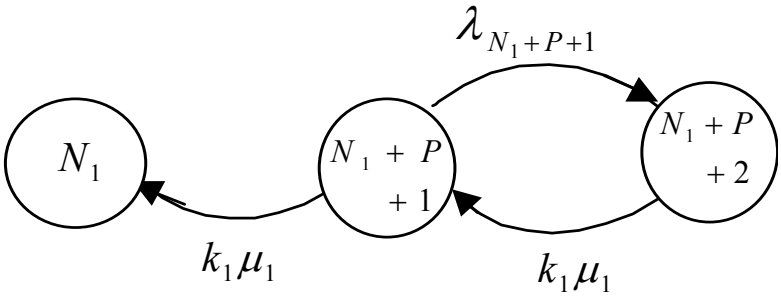

Figure 5. Process in state N_1+P+1.

4.2. Double Level Hysteretic Control In this model, it is assumed that the service time is distributed exponentially and the average rate will change between μ_1 and $k_1 \times \mu_1$.

When the queue length increases from a value less than or equal to N_1 to a value equal to or greater than N_2, then $k_1 \times \mu_1$ rate is maintained until the queue length drops to the value N_1<N_2 at which time the $k_1 \times \mu_1$ rate decreases to μ_1. Figure 3 shows a double level control. In this problem, we define N_2 - N_1=P+1.

We proceed in the analysis by first enumerating the states in the sequence $\tilde{(}0, N_1)$ with queue's lengths from 0 till N_1 and the average service rate μ_1, (N_1+1, N_1+P) with queue's lengths from N_1+1 till N_1+P and the average service rate μι, (N_1+P+1, N_1+2P+1) with queue's lengths from N_1+1 till N_1+P+1 and the average service rate k_1 × μ_1 and (N_1+2P+1, ∞) for queue's lengths greater than N_2 and the average service rate $k_1 \times \mu_1$.

4.2.1. Steady-State Equations for the States in the Sequence ($\tilde{0}$, N_1) For the states from 0 to N_1, we have the following equation.

$$P_n = \frac{\rho^n}{n!} P_0 \qquad 0 \le n \le N_1 \tag{15}$$

4.2.2. Steady-State Equations for the States in the Sequence (N_1+1, N_1+P) Figure 4 shows how to calculate $P'N_1$+1. For calculating the state $P'N_1$+P+1, we can proceed as shown in Figure 5.

$$P'_{N_1+1} = (1 + \frac{\rho}{N_1+1}) P_{N_1-1} - k_1 P'_{N_1+P+1} \tag{16}$$

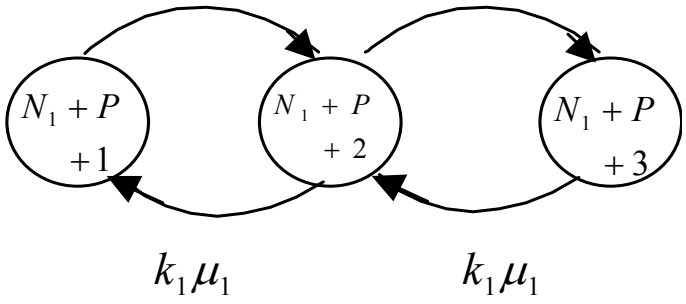

Figure 6. Process in state N_1+P+2.

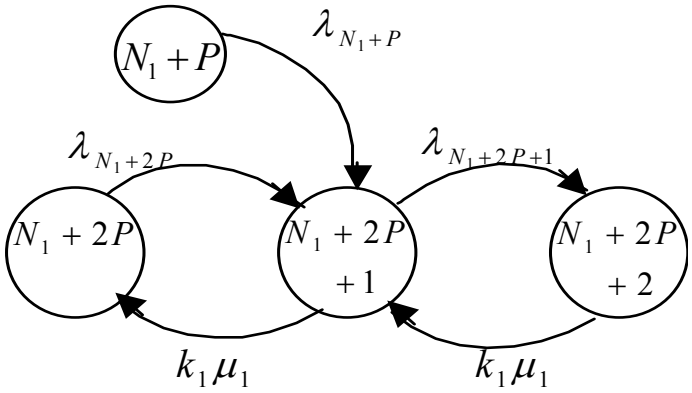

Figure 7. Process in state N_1+2P+1.

As shown in Figure 3, each of the states from N_1+2 to N_1+P is dependent on its former and latter ones. Since the rate of service is μ_1, then each state can be calculated as follows:

$$P'_{N_1+2} = (1+\frac{\rho}{N_1+2})P'_{N_1+1} - \frac{\rho}{N_1+1}P_{N_1}$$

$$P'_{N_1+P} = (1+\frac{\rho}{N_1+2})P'_{N_1+P-1} - \frac{\rho}{N_1+P-2}P_{N_1} \qquad (17)$$

$$(1+\frac{\rho}{N_1+P+1})P'_{N+P} = \frac{\rho}{N_1+P}\ P'_{N+P-1} \Rightarrow$$

$$P'_{N+P} = \frac{\frac{\rho}{N_1+P}}{1+\frac{\rho}{N_1+P+1}}P'_{N_1+P-1} \qquad (18)$$

4.2.3. Steady-State Equations for the States in the Sequence (N_1+P+1, N_1+2P+1) As shown in Figure 6, one can reach to N_1+P+1 only through the state N_1+P+2, in which the queue length is N_1+1.

$$P'_{N_1+P+1} = \frac{1}{1+\frac{\rho_1}{N_1+2}}P'_{N_1+P+2} \qquad (19)$$

For the states from N_1+P+3 to N_1+2P+1, we can

TABLE 1. Total Costs in Units of c ($\rho = 2$ and $\rho_1 = 0.5$).

	Simple $\rho = 2$	$\rho_1 = 0.5$	Single Level $N_1 = 6, N_2 = 7$	Single Level Hysteretic $N_1=0, N_2=7$	Hysteretic Double Level $N_1 = 5, N_2 = 7$	$N_1 = 3, N_2 = 7$
S_1	2.283	2.838	3.387	3.164	2.925	3.080
S_2	2.283	3.393	3.850	3.175	2.925	3.078
S_3	2.283	4.503	4.775	3.197	2.926	3.076
S_4	2.283	4.503	3.082	3.156	2.925	3.081

proceed according to the set of Equations 20, expecting the queue length from N_1+1 to N_2 with service rate $k_1 \times \mu_1$.

$$\begin{cases} P'_{N_1+P+3} = (1+\dfrac{\rho_1}{N_1+3})P'_{N_1+P+2} - \dfrac{\rho_1}{N_1+2}P'_{N_1+P+1} \\ P'_{N_1+2P+1} = (1+\dfrac{\rho_1}{N_1+P+1})P'_{N_1+2P} \\ -\dfrac{\rho_1}{N_1+2}P'_{N_1+2P-1} \end{cases} \tag{20}$$

4.2.4. Steady-State Equations for the States in the Sequence (N_1+2P +1, ∞) Figure 7 shows the state N_1+2P+1 for calculating $P'N_1$+2P+2.

$$\begin{aligned} \mathrm{P}'_{N_1+2P+2} &= (1+\frac{\rho_1}{N_1+P+2})\mathrm{P}'_{N_1+2P+1} - \\ &-\frac{\rho_1}{N_1+P+1}P'_{N_1+2\mathrm{P}} - \frac{\rho_1}{N_1+P+1}P'_{N_1+P} \end{aligned} \tag{21}$$

For calculating the probabilities from N_1+2P+1 to infinity, we can proceed similar to the Equations 15 with this difference that the queue length is from N_1+P+1 to infinity and the service rate is $k_1 \times \mu_1$.

$$P'_{N_1+2P+i+1} = \frac{\rho'_1}{l!}P'_{N_1+2P} \tag{22}$$

4.2.5. Combining the Equations to Obtain the State Probabilities Equations 19 and 20 yield the following equations.

$$\mathrm{P}'_{N_1+P+j} = \left[1+\sum_{l=1}^{j-1}\frac{{\rho_1}^l}{\prod_{i=j-l+1}^{j}(N_1+i)}\right]\mathrm{P}'_{N_1+P+1} \tag{23}$$

; $\quad 2 \le j \le P+1$

$$\begin{aligned} \mathrm{P}'_{N_1+1} &= \frac{\rho}{N_1+1}.\frac{\rho^{N_1}}{N_1!}\mathrm{P}_0 - k_1\mathrm{P}'_{N_1+P+1} \\ &= \frac{\rho^{N_1+1}}{(N_1+1)!}\mathrm{P}_0 - k_1\mathrm{P}'_{N_1+P+1} \end{aligned} \tag{24}$$

By combining Equations 15 and 17, we have the following equations.

$$P'_{N_1+i} = \frac{\rho^{N_1+i}}{(N_1+i)!}P_0 - k_1\left[1+\sum_{l=1}^{i-1}\frac{\rho^l}{\prod_{j=i-l+1}^{i}(N_1+j)}\right]P'_{N_1+P+1}$$

; $\quad 2 \le i \le P$

(25)

We can find the set of Equations 26 by combining Equations 18 and 25.

$$P'_{N_1+P+1} = \frac{\rho^{N_1+P}\rho_1}{(N_1+P+1)} \left[1+\frac{\rho}{N_1+P+1}-\frac{\rho}{N_1+P}+ \sum_{l=1}^{P-1}\frac{\rho^l}{\prod_{j=P-l+1}^{p}(N_1+j)}-\sum_{l=1}^{P-2}\frac{\rho^l}{\prod_{j=P-l}^{p-1}(N_1+j)} \right] P_0 \quad (26)$$

By summing $P_0^{'}$s coefficients and investigating them, we can calculate the value P_0. After obtaining P_0, the other state probabilities, which are necessary to compute η and F from Equations 6 and 7, can be easily obtained.

5. COST STRUCTURE

The choice of one queuing control method among several must be governed by an objective function. If a cost formula can be developed which fully accounts for the costs associated with the queuing system, then the calculated costs provide such an objective function. The total cost will be assumed to consist of a cost C_s associated with service and a cost C_q associated with the queue.

So as to limit the cost considerations to a reasonable length, no attempt will be made to derive optimum solutions because explicit optimization formulas are both difficult to obtain and become too complex for easy interpretation. Although for each particular cost formula the optimum can be obtained by developing parametric curves, the following treatment is sufficient to demonstrate that the optimum control method depends on the cost formula. The cost of service is similar to the Gebhardt's cost of service (Equation 8), with this difference that we have

$$\eta = \bar{\lambda} P'_{N_1+P} \quad (27)$$

where

$$\bar{\lambda} = \sum_{n=0}^{\infty}\lambda_n P_n = \sum_{n=0}^{\infty}\frac{\lambda}{n+1}P_n \quad (28)$$

The cost of queuing is considered as follows:

$$C_q = c_2\frac{\sum_{n=0}^{\infty} nP_n}{\lambda} = c_2\frac{\sum_{n=0}^{\infty} nP_n}{\sum\frac{\lambda}{n+1}P_n} \quad (29)$$

Cq is expressed the waiting time of a particular customer. Our main aim is to decrease the sum of the above-mentioned costs.

6. NUMERICAL EXAMPLE

For showing the numerical stability of the theoretical developments of the paper, we solve a numerical example. It is assumed that $\lambda = 2.22$ per time unit and k_1=1 or 4 according to whether one or four servers are employed. It is also assumed that the system can only operate at the average service rates of μ_1=1.11 or 444 per time unit, corresponding to ρ=2 and ρ_1=0.5, and to k_1=1 and 4. Moreover, we assume that c_1=c_2=c. As mentioned, the queuing process would be stable even without use of any control, because of considering the lapse theory in the proposed model. Taking into account the values of r_1 and r_2 in the cost formula, four versions of formula S will be used as follows:

$$\begin{cases} S_1: & r_1 = 0.5, & r_2 = 1 \\ S_2: & r_1 = 1, & r_2 = 2 \\ S_3: & r_1 = 2, & r_2 = 4 \\ S_4: & r_1 = 2, & r_2 = 0 \end{cases}$$

C_s and C_q are computed from Equations 8 and 29, respectively. Total costs (C_s+C_q) in units of c are tabulated in Table 1.

The queuing control doctrine includes the simple control at each of the two permissible service rates, the single level control, the unilevel hysteretic control ($N_1 = 0$) and the double level hysteretic control with two different settings on N_1. As observed in the table, we find the simple control with $\rho = 2$ to be best in all cases. Results obtained by double level and unilevel hysteretic models show the smaller values comparing with the single level control.

7. CONCLUSION

In this paper, two models namely single level and double level hysteretic are investigated. It was proved that the total cost of the double level hysteretic model is smaller than the single level control. Hence, its use is suggested in service-providence systems.

However, solutions obtained by the different models suggest that the simple control gives the most satisfactory results. The double level hysteretic, the unilevel hysteretic and the single level come as the second, third and fourth best, respectively.

Gebhardt [6] incorporated different coefficients in his work. He has concluded that the results obtained by the single level model are better if the cost of additional servers is smaller than the switching cost, when no lapse condition comes into action. However, if the cost of additional servers is greater than the cost of switching, then the double level model would be the best.

8. REFERENCES

1. Hiller, F. S. and Lieberman, G. J., "Introduction to Operations Research", 6th Ed., McGraw-Hill Co., NY, (1995).
2. Dilworth, J. B., "Production and Operations Management", Manufacturing and Services, 5th Ed., McGraw-Hill Co., (1993).
3. Stidham, S., "L = λW: A Discounted Analogue and A New Proof", ***Oper. Res.***, Vol. 20, (1972), 1115-1126.
4. Crabill, T. B., "A Classified Bibliography of Research on Optimal Design and Control of Queues" ***Oper. Res.***, Vol. 25, (1976), 219-232.
5. Yadin, M. and Naor, P., "On Queuing Systems with Variable Service Capacities", ***Nav. Res. Log. Quart.***, Vol. 14, (1967), 43-54.
6. Gebhardt, T., "A Queuing Process with Bi-Level Hysteretic Service Rate Control", ***Nav. Res. Log. Quart.***, Vol. 14, (1967), 55-68.
7. Sobel, M. J., "Optimal average-cost policy for a queue with start-up and shut-down costs", ***Oper. Res.***, Vol. 17, (1968), 145-162.
8. Tembe, S. and Wolff, R., "The Optimal Order of Service in Tandem Queues", ***Oper. Res.***, Vol. 24, (1974), 824-832.
9. Chao, X., "Triggered Concurrent Batch Arrivals and Batch Departures in Queuing Networks" ***Theory and Applications***, Vol. 10, (2000), 115-129.
10. Crabill, T. B., "Optimal Control of A Service Facility with Variable Exponential Service Times and Constant Arrival Rate", ***Manage. Sci.***, Vol. 18, (1972), 560-566.
12. Tijms, H., "A Control Policy for A Priority Queue with Removable Server", ***Oper. Res.***, Vol. 22, (1973), 833-837.