선전과 이탈이 있는 복수 서비스 대기행렬모형에 대한 시뮬레이션 분석*

권치명**, 김성연**, 정문상**, 황성원**

Simulation Analysis for Multiple-Server Queueing Model with Advertising and Balking

Chimyung Kwon, Sungyeon Kim, Moonsang Chung and Sungwon Whang

Abstract

The purpose of this paper is to analyse the manager's policy to maximize the profit in a multiple-server queueing facility with a limited queue capacity. We assume that the level. of advertizing effects on the arrival rate of customers to the facility. The model without 'word of mouth effect' is assumed that the arrival rate is independent on the quality of service level. We estimate the service quality by the balking rate of customers from system. We extend this to the model with 'word of mouth effect'.

To achieve the maximum profit, the most important factor is the considerably high utilization of facility for both models. Given service rate, we should maintain an effective arrival rate to some extent. To this end, among the available options, an increase of advertizing effort is more desirable than reducing the fee if the service value of customers remains unchanged.

We also investigate whether the variability of service time has a significant impact on determining the optimal policy. The cost of service variability is not so expensive as that in a single server model due to the reduced variability of service times in a multiple-server model.

Key Words; Multiple-Server Queueing Model, Word of Mouth Effect, Advertizing, Balking Rate, Optimal Policy.

* 이 논문은 2002학년도 동아대학교 학술연구조성비(연구소공모과제)에 의하여 연구되었음.

** Division of Management Information Systems, Dong-a University

1. Introduction

In this paper, we consider a multiple-server facility with limited queue capacity and a situation where a manager decides on the optimal utilization of a facility. Customers arrive at the facility to get a certain benefit from the service, but they encounter the annoyance from waiting. Arriving at the facility, they observe the service fee and the existing queue length, and decide whether or not to join. For instance, customers arrive at the game room, observe the fee and the number of players and then decide to join the queue to play.

A number of studies have analyzed control problems associated with congestion in service facilities. Naor[4] first studied the situation where arriving customers are admitted or not based on the observed queue length for a single-server facility. A similar modeling approach has been extended by various others. Stidham[6] introduced a fixed reward for each job passing through the system and a waiting cost, and studied the optimal control of admission to a queueing system. Dewan and Mendelson[2] focused on an optimal pricing and capacity decision for a service facility where user delay cost function is nonlinear. Van Ackere and Ninios[7] considered the case where the arrivalrate is determined by the level of advertising and the 'word of mouth' effect. They analysed the manager's policy maximizing the profit for a single server facility. Atkinson[1] re-analyzed Van Ackere and Ninios's model using queueing theory and numerical optimization when service time has an Erlang distribution.

This research studies the optimal strategy for a multiple-server queueing model with advertising and balking. This model is an extension of the model developed by Van Ackere and Ninios[7] to a multiple-server facility. To attract customers, manager uses advertizing, which may influence the arrival rate to a facility. Also the price of entering may determine whether or not customers join the queue. Thus, in our model, a manager's choice of advertising level and entrance fee may effect on the utilization of facility. If a manager advertises too much to attract customers, this may create congestion and many customers will balk from the system. Consequently a bad reputation will cause future advertising less effective('word of mouth' effect). Usually the effectiveness of advertizing will be evaluated by the fraction of balked customers. On the other hand, lower entrance fee may attract more customers to join the facility, so it will increase the utilization of system, but it may reduce the total profit.

In such a situation, we try to find the manager's profit maximizing policy for determining the advertising level and the price of entrance in a steady state situation. We also investigate the influence of the variability of service time on system utilization. We explore that to what extent, the manager's optimal policy depends on the variation of service time. For these purpose, we developed simulation models to study optimal policies which determine advertizing expenditures and fees in various cases of service times.

2. Model with Balking and Advertising

We suppose that customers arrive at the system according to a Poisson process with arrival rate λ. The manager can decide an expenditure of advertising which influences the arrival rate. We consider two cases: (a) in the first case, the arrival rate is only determined by an advertizing expenditure, A and we assume that the arrival rate is a linear function of advertizing, $\lambda = cA$(c is a constant), (b) secondly, we extend the first case to a model incorporating word of mouth effect. For a given fee F, high level of advertizing induces more customers to arrive at the system, which will result in the increased fraction of balking customers from system. It finally causes a bad reputation for the facility and makes future advertizing less effective. The effectiveness of advertizing level is assumed to be measured by a function of service quality. Specifically, in the second case, we assume that the effective arrival rate is reduced by the balking probability of customers from system.

When customers arrive at the system, they observe the queue length, and entrance fee and then decide whether or not to participate in queue for services. We assume that a service is worth a value, V for all customers. We also assume that the waiting cost of a customer is linear in the waiting time, W. We consider that the facility has m identical servers whose mean service rates are all equal to μ. To standardize waiting cost per time unit in a similar way of Van Ackere and Ninios [7], we let the waiting cost per time unit be equal to μ. Then a customer's valuation of the service is equal to $V-F-\mu W$. If an arriving customer perceives k persons including himself in the system, he might expect to spend $k/m\mu$ time units in the system. Therefore he will join the system if

$$V-F-\mu k/m\mu = V-F-k/m \geq 0. \quad (1)$$

This implies that the manager's choice of F implicitly determines the maximum queue length, $K=m(V-F)$ $(K \geq 1)$. Thus, for given V and F, an arriving customer will join the queue if he observe the queue length less than $m(V-F)$. Otherwise he turns away from the system.

The system manager tries to determine the variables A and F which yield the maximum profit per time unit. We let P_K denote the balking probability of customers from system when arriving customers find K customers in the system. For a model without 'word of mouth effect', the profit function per unit time is given by as follows:

$$Profit(1) = \lambda(1-P_K)F-A, \quad (2)$$

where the first term means the earning from customers who join the queue, which is the arrival rate times fraction of customers served times fee per customer; and the second term simply denotes the advertizing expenditure. We note that the P_K is a non-linear function of A and F. Plugging the relationship $\lambda = cA$ into (2) yields

$$Profit(1) = \lambda(1-P_K)F-\lambda/c, \quad (3)$$

where c is a constant implying the coefficient of advertizing expenditure on arrival rate. Next we consider the profit function for the model including the word of mouth effect. In this case, the arrival rate is reduced by the amount of P_K, which is due to the bad reputation from customers balking from the system. We assume that the effective arrival rate to the facility is equal to $\lambda = cA(1-P_K)$. Similarly in developing the equation (2), we present the objective function as follows:

$$\begin{aligned} Profit(2) &= \lambda(1-P_K)F-A \\ &= c(1-P_K)^2AF-A. \end{aligned} \quad (4)$$

If we substitute for A into the above equation, then we have the equivalent profit function as follows:

$$Profit(2) = [\lambda(1-P_K)F-\lambda/(1-P_K)]/c. \quad (5)$$

For exponential service time, we can find the balking probability based on queueing model M/M/m/K. Using this probability, we can obtain the manager's optimal policy numerically. However, in cases of general service-time distributions, since obtaining the balking probability analytically is not available, analyzing this system through simulation is recommended.

3. Simulation Model

We develop a simulation model to study the manager's policy for determining the advertizing level and service price in cases of non-exponential service times. Since we have some difficulties in analytically obtaining the balking rates of customers from system for non-exponential service time, we estimate them through simulation. To validate a corresponding simulation model against an analytical model, we simulate the model with the exponential service time, and we compare simulation results with analytical ones. To this end, we build the simulation model of the multiple-server queueing system with limited capacity by using simulation package SLAM II[5]. We first present the analytical and simulated results for exponential service times. Based on these results, we then conduct a set of simulation runs to investigate the effect of variability of service-time on the profit functions for cases that service times have the general distributions.

3.1 Simulation Model with Exponential Service Time

We consider the multiple-server queueing model with number of identical servers $m=4$ and entrance fee $F=8$. A service time of each server is exponentially distributed with mean service rate $\mu=1$. To all customers, the service from system is worth of the value $V=10$ and they arrive at the system according to a Poisson process with arrival rate λ. For given $V=10$, $F=8$ and $m=4$, we see that the maximum capacity of queue is equal to $K=8$. Based on queueing theory [3], we can analytically calculate the balking probability of customers, P_K as follows:

$$P_k = P_0(\lambda/\mu)^k 1/m!(1/m)^{k-m}, \quad (6)$$

where

$$P_0 = [\sum_{k=0}^{m} (\lambda/\mu)^k/k! + \sum_{k=m+1}^{K} (\lambda/\mu)^k m^{m-k}/m!]^{-1}$$

For 10 different values of λ(=1, 2, ..., 10), we compute the P_K by the equation (6).

Under the same conditions as given in the analytical model, through the simulation runs, we estimate the P_K as the ratio of number of customers balked to sum of number of customers both balked and serviced. We stop simulation runs when number of customers serviced reaches 10,000. Figure 1 graphically shows the analytical and simulated results of the balking rate P_K for $\mu=1$, and an arrival rate λ ranging from 1 to 10 by increasement of 1.

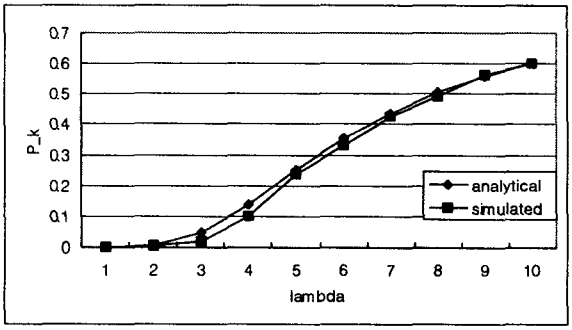

Figure 1. Analytical and Simulated Values of P_K

The balking rate seems to be a little underestimated in the range of λ from 1 to 8. As we see in Figure 1, the lower λ tends to show the worse estimation. This result is similar to that obtained by Van Ackere and Ninios for a single-server case [1]. As they noted, this discrepancy is not a serious problem because (a) the large percentage error has little meaning as the true values of P_K are close to zero, and (b) the profit function is maximized for a not small values of utilization factor $\rho=\lambda/m\mu$.

Next we compute two profit functions, *Profit*(1) and *Profit*(2) by use of analytical method as well as simulation runs. We may assign different values for c in (3) and (5), but simply, we set c equal to 1 as Van Ackere and Ninios's model since this research focuses on extending of a single-server model to a multiple-server case, and investigating shapes of profit functions. In estimating the unit time profit functions (3) and (5) through simulation, we calculate them as follows:

$$Profit(1) = \frac{\text{no of customers serviced}}{\text{simulation duration}} F - \frac{\text{total no of customers}}{\text{simulation duration}}, \quad (7)$$

$$Profit(2) = \frac{\text{no of customers serviced}}{\text{simulatin duration}} F - \frac{\text{total no of customers}}{\text{simulatin duration}} \frac{1}{1-P_K} \quad (8)$$

Figure 2 and 3 (4 and 5) present the analytical and simulated profits of model without (with) the effect of word of mouth for different values of the fee F, respectively.

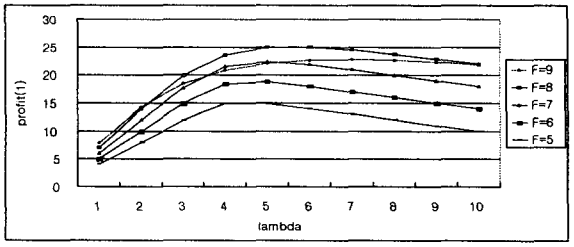

Figure 2. Analytical *Profit*(1) for several values of F

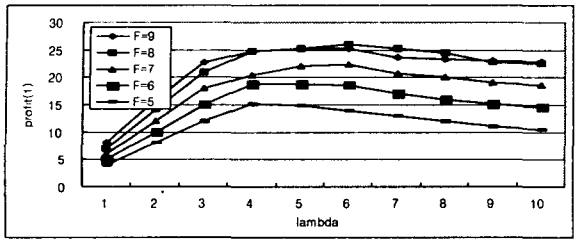

Figure 3. Simulated *Profit*(1) for several values of F

From Figures 2 and 3, we see that analytical and simulated profit functions are similar, and both functions have the same optimal policies: $\lambda=6$, $F=8$. However, the optimal values of simulated profit function is a little higher than that obtained by the analytical method. We consider that this is due to lower estimation of P_K at $\lambda=6$ (see Figure 1).

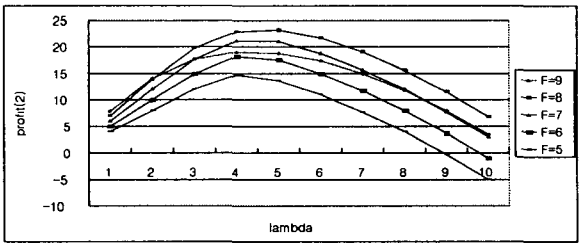

Figure 4. Analytical *Profit*(2) for several values of F

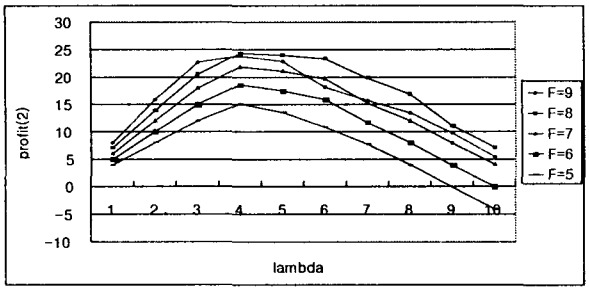

Figure 5. Simulated *Profit*(2) for several values of F

For the model with the word of mouth effect, two profit functions also show similar trends and their values are lower than those without the word of mouth effect. Figure 4 and 5 show that, as the arrival rate increases, the increased balking rate gives customers a bad reputation, and decrease profit substantially in case that λ is higher than 8. The optimal policies of analytical and simulation method are $F=8$ and $\lambda=5$, and $F=8$ and $\lambda=4$, respectively. The profits from simulation run is a little higher than those of analytical method. Similarly to the previous case, we consider this is because of underestimation of P_K in simulation run.

3.2 Experiment on Simulation Model with General Service Time

We conducted a large set of simulation experiments on a queueing model M/G/m/K with number of indertical servsers, $m=4$, to evaluate the manager's policy for determining the advertizing level and service fee for general service-time distributions, and to investigate their effect on the profit. Specifically we consider 4 different distributions of service times, S whose expectations and variances are follows:

(1) Deterministic: $E(S)=2$, and $Var(S)=0$,
(2) Erlang: $E(S)=2$, and $Var(S)=2$,
(3) Gamma: $E(S)=2$, and $Var(S)=4$.
(4) Gamma: $E(S)=2$, and $Var(S)=8$.

Under the condition that $V=10$ and $F=9$, Figure 6 and 7 present the simulated profits based on 4 different service-time

distributions for the model without and with word of mouth effect, respectively. Table 1 and 2 summarize the optimal policies and operating characteristics of models with and without word of mouth effect for considered distributions. In Table 2, the value of advertizing expenditure A is computed by the equation of $A=\lambda/(1-P_K)$.

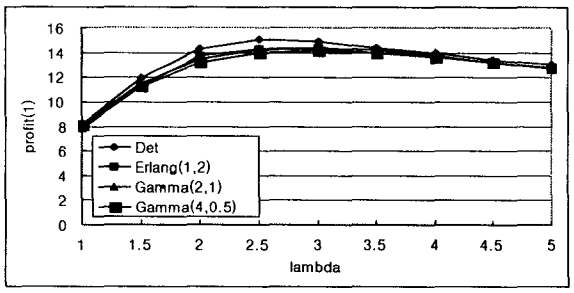

Figure 6. Simulated *Profit*(1) for 4 service-time distributions.

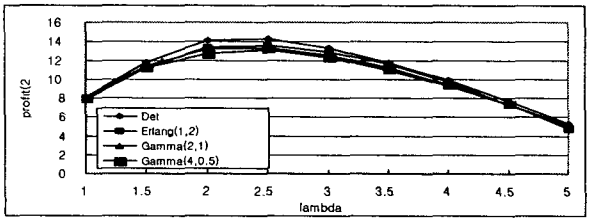

Figure 7. Simulated *Profit*(2) for 4 service-time distributions.

Table 1. Optimal policies for the model without word of mouth effect

Distribution	λ	P_K	F	Waiting Time	Queue Length	Profit
Deterministic	2.5	0.229	9	1.188	2.326	15.1
Erlang(1, 2)	3	0.347	9	1.382	2.674	14.4
Gamma(2, 1)	3	0.367	9	1.350	2.596	14.3
Gamma(4, 0.25)	3	0.372	9	1.270	2.423	14.1

Table 2. Optimal policies for the model with word of mouth effect

Distribution	λ	P_K	A	F	Waiting Time	Queue Length	Profit
Deterministic	2.5	0.229	3.24	9	1.188	2.326	14.3
Erlang(1, 2)	2.5	0.248	3.32	9	1.152	2.148	13.5
Gamma(2, 1)	2.5	0.266	3.41	9	1.135	2.109	13.3
Gamma(4, 0.25)	2.5	0.274	3.44	9	1.091	2.005	13.1

4. Discussion of Simulation Results

We first provide a summary of simulation results for the model without word of mouth effect. The optimal policy for deterministic service time is given by $F=9$ and $\lambda=2.5$, and the optimal policies for Erlang(1, 2), Gamma(2, 1) and Gamma(4, 0.25) service times are all given by $F=9$ and $\lambda=3$. The values of profit functions are in the range from 14.1 to 15.1. As we expected, the larger the variance of service time is, the more customers balk, and the less profit we have. In a deterministic service time case, a less advertising level achieves more profit than profits obtained from three other service times. We see that the variance of Gamma(4, 0.25) is two times and four times of that of Gamma(2,1) and Erlang(1,2), respectively, but the differences between two optimal profits are not much(0.3 and 0.2, respectively). For a multiple-server system, variance is not so serious as that in a single-server system. We consider that this is due to the reduced variability of service distribution for multiple servers.

In case of Gamma(4, 0.25) service distribution, if we increase the advertising

level by the amount 0.5 (from 3 to 3.5), for instance, to compensate the service variability, the profit decreases from 14.1 to 14.0. Thus the advertising effort to increase the arrival rate is not always desirable even though the variability of service time is large.

Given the fee and advertising level, as the variability of service time increases, the balking rate increases and thus, effective arrival rate decreases. Therefore an increase in advertising to boost arrival to some extent seems desirable. In our example, compared to service rate $\mu(=0.5)$, the optimal arrival rate λ is considerably high. We note that the system utilization factor $\rho=\lambda/m\mu=1.5$ for Erlang and Gamma distributions. If we increase the arrival rate from 3 to 4, the balking rate of customers from system increases from 0.372 to 0.513 in Gamma(4, 0.5) case, and hence the profit drops from 14.1 to 13.6. Similar results are observed for other cases.

Secondly we discuss simulation results for the model including word of mouth effect. The optimal policies for all cases are given by $F=9$ and $\lambda=2.5$. Compared with model without word of mouth effect, the optimal policy for the deterministic case is same, but the optimal policies for three other distributions are given by the same fee $F=9$, and a little less arrival rate $\lambda=2.5$. We consider that this result is reasonable because of word of mouth effect. Except deterministic service time case, changing λ from 3 to 2.5 shows the decrease of balking rate of around 10%, and thus increasing balking rate causes advertising to attract customers less effective.

Throughout two models with four different service distributions, we have the optimal policies given by $F=9$ and λ in the range from 2.5 to 3. We have several available options to achieve the maximum profits: (a) change the advertising level, (b) change the fee, and (c) use a combination of (a) and (b). Simulation results from our example suggest that a considerable advertising effort to boost customers is an important factor to achieve the maximum profit. Even in the model allowing the word of mouth effect, the utilization factor is 1.25. We note a similar result given in a single server model [7]. For maintaining an arrival rate to some extent, we may increase the advertising expenditure, reduce the fee, or use both options. If the service value of customers remains unchanged, an option of reducing the fee is not desirable. Reduced fee induces more customers to join the system (increased queue length and waiting time). But it also results in more congestion of system and reduces the earning from each customer. Totally, an effect of reduced fee to the profit is not much in our example.

Finally we note that the profit decreases as the variance of service times increases for both models. This is the same result as obtained from a single server system. But the cost of service variability is not so expensive as that in a single server model. We conjecture that this is due to the reduced variability of service times when multiple servers serve the customers.

5. Conclusions

Our purpose is to analyse the manager's policy to maximize the profit in a multiple

-server queueing facility with a limited queue capacity. In determining the optimal policy, the most important factor is the considerably high utilization of facility. Hence given service rate, we should maintain an effective arrival rate to some extent. Among the available options for this, an increase of advertizing effort is more desirable than reducing the fee if the service value of customers remains unchanged.

For the model with word of mouth effect, an increased balking rate causes to drop the future arrival rate of customers, and it will reduce the effectiveness of advertising. Even in this case, advertising to boost arrival to some extent seems desirable. In order to have any concrete idea on the optimal utilization factor, we need a large set of simulation experiments on various systems. Our simulation results and study of Van Ackere and Ninios suggest the optimal value of utilization may be in the range from 1 to 1.5. in many real systems. We also observe that the variability of service time has an impact on determining the optimal policy. However, the cost of service variability is not so expensive as that in a single server model. We consider this is due to the reduced variability of service times in a multiple-server model.

Despite of limited simulation runs on specific models, we expect this study may be useful in determining the optimal policy for a multiple-server facility with advertizing and balking.

Acknowledgment: The authors would like to thank the referees for their helpful comments.

References

[1] Atkinson, J. B. (1996). A Note on a Queueing Optimization Problem, *Journal of the Operational Research Society* 47, 463-467.

[2] Dewan, S. and Mendelson, H. (1990). User Delay Costs and Internal Pricing for a Service Facility, *Management Science* 16, 1502-1517.

[3] Kleinrock, L. (1975). *Queueing System Vol I: Theory*, John Wiley & Sons, New York.

[4] Naor, P. (1969). On Deregulation of Queueing Size by Levying Tolls, *Econometrica* 37, 15-24.

[5] Pritsker, A.A.B. and O'Reilly, J. (1999). *Simulation with Visual SLAM and AweSim*. John Wiley & Sons, New York.

[6] Stidam, S. (1985). Optimal Control of Admission to a Queueing System, *IEEE Transaction on Automatic Control* AC-30, 705-713.

[7] Van Ackere, A. and Nininos, P. (1993). Simulation and Queueing Theory Applied to a Single-server Queue with Advertising and Balking, *Journal of the Operational Research Society* 44, 407-414.

주 작 성 자 : 권 치 명
논문투고일 : 2003. 09. 16
논문심사일 : 2003. 10. 02(1차), 2003. 11. 04(2차), 2003. 11. 18(3차)
심사판정일 : 2003. 11. 20

● 저자소개 ●

권치명

1978년 서울대학교 산업공학과 졸업
1983년 서울대학교 대학원 산업공학과 졸업
1991년 VPI & SU 산업공학과 박사
현 재 동아대학교 경영정보과학부 교수
관심분야 : Simulation Modeling & Output Analysis, Simulation Optimization, FMS

김성연

1981 서울대학교 계산통계학과 (이학사)
1983 서울대학교 대학원 통계학전공 (이학석사)
1997 North Carolina State University (통계학박사)
현재 동아대학교 경영정보과학부 교수
관심분야 : 선형모형, 비선형모형, 혼합모형

정문상

1979 서울대학교 경영대학 학사
1981 한국과학기술원 경영정보시스템 전공 (석사)
1988 한국과학기술원 경영정보시스템 전공 (박사)
현재 동아대학교 경영정보과학부 교수
관심분야: MIS, 시스템 분석 및 설계, 정보 통합

황성원

1987 동아대학교 응용통계학 학사
1993 경성대학대학원 정보공학 석사
2003 동아대학교 경영정보과학부 박사과정 수료
현재 동의대학교 겸임교수
관심분야: MIS, 시스템 분석 및 설계, 정보 통합