Optimal Control in a Single Server Queueing System with Setup Times

Yan Ma

***Abstract*—In this paper, we study the optimal control of an M/M/1 queue with exponential setup times. Customer's equilibrium joining strategies and facility manager's profit-maximizing pricing policies are explored under two common toll structures. The first one is flat fee model and the second one is time-based fee model. It is found the two different charging mechanisms do not affect customers' equilibrium joining rate, server's maximal profit and social welfare. Furthermore, the price decision of a profit-maximizing server is always socially optimal. In conclusion, under customers' equilibrium, the two different toll schemes are equivalent from the economic point of view.**

***Index Terms*—Queueing system, Setup times, Optimal control, Nash equilibrium, Profit maximizing.**

I. Introduction

IN many situations, controlling quality of service is an important task. Indeed, when resource is limited and demand is high, congestion occurs and service completion delay may increase to an unacceptable level. By choosing a suitable pricing scheme for the real conditions, facility managers can control congestion and allocate resource properly. A wise decision can benefit both the toll collectors and customers to get to a win-win status.

Tolling a flat fee or charging by time are two common pricing schemes in real life. It is expected that the two different pricing policies may lead to different customer's entering rate as well as different server's profits. Hence, how to choose pricing schemes so that the service providers can make a maximal profit and at the same time customers can achieve Nash equilibrium state (no one can benefit by changing the equilibrium strategies) is a valuable subject to explore.

Generally, a service system can be described as a queueing model. The queueing theory has been rapidly developed in the past several decades. A lot of research has been carried, such as [13], [9], [10]. In many queueing systems, a server may be deactivated for economic reasons, suffer random failures, go under preventive maintenance or attend to a secondary system. Such situations often incur in real applications. Due to the versatility and applicability, queueing systems with removable servers have been extensively explored. Detailed surveys are contained in [14,15]. Especially, a significant portion of these literatures are dedicated to queues with setup times. In such models once a server is reactivated, a random time is required for setup before it can begin serving customers. Initially, Doshi [5] explored a GI/G/1 queueing system with setup times. Krishna Reddy et al. [6] examined a bulk queueing model with multiple vacations and setup times. Then Choudhury [4], Bischof [1] and the references therein analyzed various single-server systems with setup times. Recently, Yajima and Phung-Duc [17] dealt with a Batch arrival single-server queue with setup times.

Meanwhile, with the development of economics of queueing systems, optimal pricing issues and equilibrium control in service facilities have attracted more and more attention. Low [7] considered an optimal dynamic pricing policies in an M/M/s queue where customers' arrival rate is a strictly decreasing function of currently advertised price. Mendelson and Whang [8] studied optimal incentive-compatible priority pricing for an M/M/1 queue. Stidham [11] analyzed optimal pricing and capacity for a service facility, in which the design variables are the service rate and the arrival rate. Wang et al. [16] presented the cost analysis of a Discrete-time queue. Subsequently, Chen and Frank [2] carried out research that allows the firm to adjust price to the state of demand for maximizing social welfare. Subsequently, Sun and Li [12] concerned the effect of information and pricing strategies on servers profits in an unobservable single server queue. Chen and Zhou [3] focused on the equilibrium strategies in the single server queue with setup times and breakdowns. Zhang and Wang [18] studied equilibrium strategies in an M/G/1 retrial queue with reserved idle times and setup times.

In the present paper, we study equilibrium strategies for customers and optimal pricing polices for facility managers. On one hand, Nash equilibrium state is favorable among customers since no one will benefit by changing it. To identify equilibrium strategies is a main objective for customers. On the other hand, facility managers concern much about toll mechanisms because it is expected they directly affect business profits. Combining the two stakeholders' goals, we investigate customers' equilibrium strategies and servers' optimal pricing polices in an M/M/1 queue with setup times. We give a multi-discussion and comparison under two common toll structures. The first one is flat fee model, in which customers are charged by a fixed fee if they decide to join the queue. It is a pretty simple pricing scheme for servers to collect fees. The second one is time-based fee model where the server tolls a fee that is proportional to the time of facility use. Thus the longer service time needed, the more fees will pay.

The reminder of this paper is organized as follows. Section 2 presents the description of the model. Section 3 develops the flat fee model. The time-based fee model is analyzed in Section 4. We derive the equilibrium joining rate and investigate profit-maximizing server's pricing behaviors. Then in Section 5, we make a comparison between the two pricing models and carry out numerical experiments. Finally, the

Manuscript received Dec. 26, 2017; revised May 21, 2018. This work was supported by the National Science Foundation of China (No.11601489), Startup Research Fund of Zhengzhou University (No.129-51090091), Outstanding Young Talent Research Fund of Zhengzhou University (No.129-32210453).

Yan Ma is with the School of Mathematics and Statistics, Zhengzhou University, Zhengzhou, P.R.China (e-mail: yan_ellen_ma@126.com).

conclusions come in Section 6.

II. Model description

Generally, facility managers concern much about the profits and want to gain the maximum value. Thus here we consider a profit-maximizing server, who sets prices to maximize his own benefit. Arriving customers do not know the queue length and server state. The decision to join or balk is irrevocable. After joining the queue, customers are served according to first-come-first-served (FCFS) discipline.

Our model is based on the following assumptions.

1. Customers arrive according to a Poisson stream with rate Λ at a single-server station. Λ is the potential arrival rate which is not necessarily the actual joining rate of customers to be served (which we refer to as λ) due to balking.
2. Service times are independently and identically distributed (i.i.d.) exponential random variables with rate μ. For system stability, we assume $\mu > \lambda$.
3. In a queue with setup times: the server is deactivated as soon as the queue becomes empty. When a new customer arrives at an empty system, a setup process starts for the server to be reactivated. The time required for setup is also exponentially distributed with rate θ.
4. Inter-arrival times, service times, and setup times are mutually independent.
5. Every joining customer receives a reward of R for completing service. There is a waiting cost of C per time unit for a customer staying in the system (in queue or in service).
6. Customers are risk neutral.
7. For the model to be non-trivial, the condition of $R \geq C\left(\frac{1}{\mu}+\frac{1}{\theta}\right)$ is assumed. This condition ensures that the reward for service exceeds the expected waiting cost for a customer joining an empty system. Otherwise, after the system becomes empty for the first time, no customers will ever enter.

We use the following notations:

ω	Expected waiting time of a customer in the system (including the service time).
P	Price charged by the service provider $(0 < P < R)$.
U	Expected utility of a joining customer.
B	Expected benefit per time unit for the service provider.
SW	Expected social welfare per time unit in the system.

Customers make joining or balking decision upon their arrival instants. There exist two kinds of possible equilibria in the system. In the first case, all the customers seeking service enter the queue, i.e. $\lambda = \Lambda$. Clearly, customers have a nonnegative expected utility so that all prefer to join in this case. Hence, the profit-maximizing server could increase his price without causing any balking until the expected customer utility becomes zero. In the second case, arriving customers follow a mixed strategy, where they join the queue with a probability such that $\lambda < \Lambda$, for which the expected customer utility is zero. Thus, there is no incentive for customers to change their joining or balking behavior.

It is possible that the customer self-interest equilibrium (denoted as λ_f in flat fee model and λ_t in time-based fee model) cannot be reached because of the arrival rate's upper bound Λ. Hence, we study two cases: unbounded case ($\lambda_f \leq \Lambda$ or $\lambda_t \leq \Lambda$) and bounded case ($\lambda_f > \Lambda$ or $\lambda_t > \Lambda$).

III. Flat fee model

Firstly, we consider the flat fee model, in which the service provider charges a fixed price. We add a subscript f at the bottom right corner of some variables. Thus, the expected utility for a customer $U = R - P_f - C\omega$.

In M/M/1 queue with exponential setup times, the mean sojourn time for a customer $\omega = \frac{1}{\mu-\lambda} + \frac{1}{\theta}$.
Suppose $\lambda_f = \mu - \frac{C\mu\theta}{\sqrt{CR\mu\theta^2 - C^2\mu\theta}}$, then we give the following theorem.

Theorem 1
(1) if $\lambda_f \leq \Lambda$, there exists a unique equilibrium where $\lambda = \lambda_f$ and $P_f = R - \frac{\sqrt{CR\mu\theta^2 - C^2\mu\theta} + C\mu}{\mu\theta}$, $B_f = \left(\mu - \frac{C\mu\theta}{\sqrt{CR\mu\theta^2 - C^2\mu\theta}}\right)\left(R - \frac{\sqrt{CR\mu\theta^2 - C^2\mu\theta} + C\mu}{\mu\theta}\right)$.
(2) if $\lambda_f > \Lambda$, there exists a unique equilibrium where $\lambda = \Lambda$ and $P_f = R - C\left(\frac{1}{\mu-\Lambda} + \frac{1}{\theta}\right)$, $B_f = \Lambda\left(R - C\left(\frac{1}{\mu-\Lambda} + \frac{1}{\theta}\right)\right)$.

Proof: (I) We begin with the case of $\lambda_f \leq \Lambda$ (Λ is large enough).

Based on the fact that in equilibrium the expected utility for a customer is zero, we obtain $R = P_f + C\left(\frac{1}{\mu-\lambda} + \frac{1}{\theta}\right)$. Hence the joining rate:

$$\lambda = \mu - \frac{C\theta}{(R - P_f)\theta - C}. \tag{1}$$

Substituting (1) into $B_f = \lambda P_f$, we have

$$B_f = P_f\left(u + \frac{C\theta}{C + \theta(P_f - R)}\right).$$

Since $R > P_f + C\frac{1}{\theta}$, the second-order derivatives of the profit function

$$\frac{\partial^2 B_f}{\partial P_f^2} = \frac{2C\theta^2(R\theta - C)}{(C + P_f\theta - R\theta)^3} < 0,$$

which means B_f is concave in P_f. Based on the necessary conditions of maximizing B_f with respect to P_f, we obtain the optimal prices of the service provider as

$$P_f = R - \frac{\sqrt{CR\mu\theta^2 - C^2\mu\theta} + C\mu}{\mu\theta}. \tag{2}$$

Substituting (2) into (1), we obtain

$$\lambda = \mu - \frac{C\mu\theta}{\sqrt{CR\mu\theta^2 - C^2\mu\theta}} = \lambda_f \tag{3}$$

Hence,

$$\omega_f = \frac{\sqrt{CR\mu\theta^2 - C^2\mu\theta}}{C\mu\theta} + \frac{1}{\theta}, \tag{4}$$

$$B_f = \left(\mu - \frac{C\mu\theta}{\sqrt{CR\mu\theta^2 - C^2\mu\theta}}\right) * \left(R - \frac{\sqrt{CR\mu\theta^2 - C^2\mu\theta} + C\mu}{\mu\theta}\right). \tag{5}$$

(II) Next we consider the case where Λ poses an effective limitation, namely $\lambda_f > \Lambda$. In this case, the profit maximization equilibrium determined by (3) cannot be reached since the maximal joining rate is Λ. Moreover, the server could increase his price from it in (2) without affecting the joining rate of the system until U equals zero. So we conclude

$$P_f = R - C\left(\frac{1}{\mu-\Lambda} + \frac{1}{\theta}\right), \tag{6}$$

$$B_f = \Lambda\left(R - C\left(\frac{1}{\mu-\Lambda} + \frac{1}{\theta}\right)\right). \tag{7}$$

IV. Time-based fee model

Next, we consider the time-based fee model, in which the service provider charges a price that is proportional to the time interval of facility use. We add a subscript t at bottom right corner of some variables. Thus, the expected utility for a customer $U = R - P_t/\mu - C\omega$.

Letting $\lambda_t = \mu - \sqrt{\frac{C\mu\theta}{R\theta - C}}$, we can give the following theorem.

Theorem 2
(1) if $\lambda_t \le \Lambda$, there exists a unique equilibrium where $\lambda = \lambda_t$ and $P_t = R\mu - \frac{\sqrt{C\mu\theta(R\theta-C)}+C\mu}{\theta}$,
$B_t = \left(\mu - \sqrt{\frac{C\mu\theta}{R\theta-C}}\right)\left(R - \frac{C}{\theta} - \sqrt{\frac{C(R\theta-C)}{\mu\theta}}\right)$.
(2) if $\lambda_t > \Lambda$, there exists a unique equilibrium where $\lambda = \Lambda$ and $P_t = \mu\left(R - \frac{C}{\mu-\Lambda} - \frac{C}{\theta}\right)$,
$B_t = \Lambda\left(R - \frac{C}{\mu-\Lambda} - \frac{C}{\theta}\right)$.

Proof: (I) Firstly, we consider the unbounded case $\lambda_t \le \Lambda$.

Based on the fact that in equilibrium the expected utility for a customer is zero, we obtain $R = P_t/\mu + C\left(\frac{1}{\mu-\lambda} + \frac{1}{\theta}\right)$. Hence the joining rate:

$$\lambda = \mu - \frac{C\theta}{(R - P_t/\mu)\theta - C}. \tag{8}$$

Substituting (8) into $B_t = \lambda P_t/\mu$, we have

$$B_t = P_t + \frac{P_t C\theta}{P_t\theta + C\mu - R\mu\theta}.$$

Since $R > P_t/\mu + C/\theta$, the second-order derivatives of the profit function

$$\frac{\partial^2 B_t}{\partial P_t^2} = -\frac{2C\mu\theta^2(R\theta - C)}{(R\mu\theta - C\mu - P_t\theta)^3} < 0,$$

which means B_t is concave in P_t. We could maximize B_t with respect to P_t, and get the optimal price as

$$P_t = R\mu - \frac{\sqrt{C\mu\theta(R\theta - C)} + C\mu}{\theta}. \tag{9}$$

Substituting (9) into (8), we obtain

$$\lambda = \mu - \sqrt{\frac{C\mu\theta}{R\theta - C}} = \lambda_t. \tag{10}$$

Hence,

$$\omega_t = \frac{1}{\theta} + \sqrt{\frac{R\theta - C}{C\mu\theta}}, \tag{11}$$

and

$$B_t = \left(\mu - \sqrt{\frac{C\mu\theta}{R\theta - C}}\right)\left(R - \frac{C}{\theta} - \sqrt{\frac{C(R\theta - C)}{\mu\theta}}\right). \tag{12}$$

(II) Next we consider the case of $\lambda_t > \Lambda$. In this case, the profit maximization equilibrium determined by (10) cannot be reached since the maximal joining rate is Λ. Moreover, the server could increase his price without causing any balking until U equals zero. So we conclude

$$P_t = \mu\left(R - \frac{C}{\mu-\Lambda} - \frac{C}{\theta}\right), \tag{13}$$

$$B_t = \Lambda\left(R - \frac{C}{\mu-\Lambda} - \frac{C}{\theta}\right). \tag{14}$$

V. Analysis and Numerical Experiments

Based on the above results, we make some conclusions and then carry out numerical experiments in this section.

In the unbounded case, we obtain the equilibrium joining rate λ_f in flat fee model and λ_t in time-based fee model. In the bounded case, Λ is always the customers' equilibrium joining rate, no matter in flat toll model or in time-based toll model. We can give the following corollaries.

Corollary 3
The charging mechanism does not affect customers' equilibrium joining rate and social welfare.

Proof: In the unbounded case, from (3) and (10), we find $\lambda_f = \lambda_t$. In the bounded case, the equilibrium joining rate is always Λ. Those results indicate the equilibrium joining rate is identical in the two different toll mechanisms. Hence, the expected waiting time ω is also the same (i.e. $\omega_f = \omega_t$). Since $SW = \lambda(R - C\omega)$, it is clear that the social welfare is thus identical in the two toll models.

Therefore, the charging mechanism makes no difference in customers' joining behaviors as well as social welfare between the flat fee model and the time-based fee model.

Corollary 4
The price decision of a profit-maximizing server is socially optimal.

Proof: In both the two charging models, the social welfare $SW = \lambda(R - C\omega)$.

In the fixed fee model, the server's profit function $B_f = \lambda_f P_f$. The customers' total expected utility $\lambda_f U = \lambda_f(R - P_f - C\omega_f)$. When the joining rate is not constrained, this expected utility equals zero in customers' equilibrium, which implies

$$B_f = \lambda_f P_f = \lambda_f(R - C\omega_f) = SW_f,$$

so that the server's profit-maximizing price decision is consistent with the society to achieve social optimization.

In time-based fee model, $B_t = \lambda_t P_t/\mu$ and $U = R - P_t/\mu - C\omega_t$. Similarly, We could conclude $B_t = SW_t$. The price chosen by profit-maximizing server also maximize the social welfare.

Corollary 5

In customers' equilibrium, the server's maximal profit in flat fee model is the same as in time-based fee model.

Proof: In unbounded case, from Corr. 3 and Corr. 4, we obtain

$$B_f = SW_f = SW_t = B_t.$$

In bounded case, from (7) and (14), we make it directly.

Remark Corr.3-5 indicate in customers' equilibrium, the two different toll mechanisms are equivalent from the economic point of view. Although customers charged by different toll systems, they enter with the same probability and the social as a whole obtain the same welfare. It is an interesting conclusion that reveals some features not being found before. Since obtaining the same maximal profits, the facility manager can simply set a fixed price instead of tolling by time so as to reduce workload.

Now we present a set of numerical experiments in the unbounded case. Here we concern about the sensitivity of customers' equilibrium joining rate λ and server's expected benefit B (also seen as social welfare SW) with respect to system parameters.

In Fig. 1 and 3, it is obvious that λ and B are increasing with θ, while decreasing with C. This means a small mean setup time or per time unit cost is beneficial for customers' joining rate and social welfare. Furthermore, if there is a smaller value for C, the sensitivity of λ or B with respect to θ turns to be less. In Fig. 2, the value of entering rate presents as a plane with the variation of μ and R. If we choose a certain value for R, there are more customers entering the queue when higher service rate is provided. However, if we choose a fixed value for μ, the joining rate varies as a slight growth line, and is not affected very much by the variation of service reward R. In Fig. 4, it is found that there is a ridge on the protruding surface, which implies a higher service rate or service reward is better for both facility managers and customers.

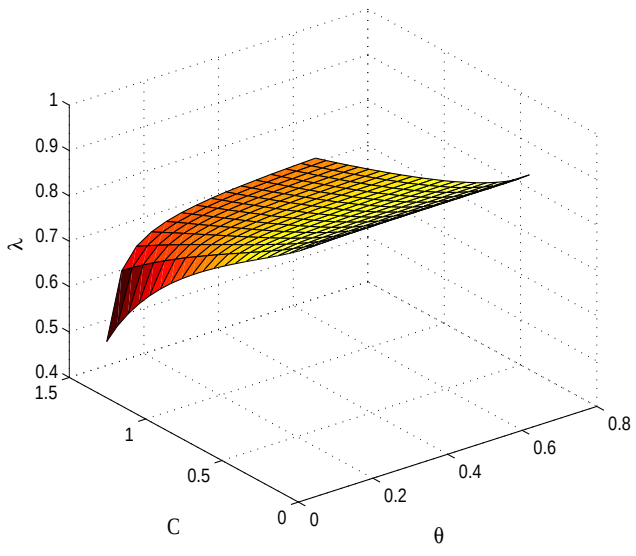

Fig. 1. Joint effect of vacation rate and per time waiting cost on λ, when $\mu = 1, R = 20$.

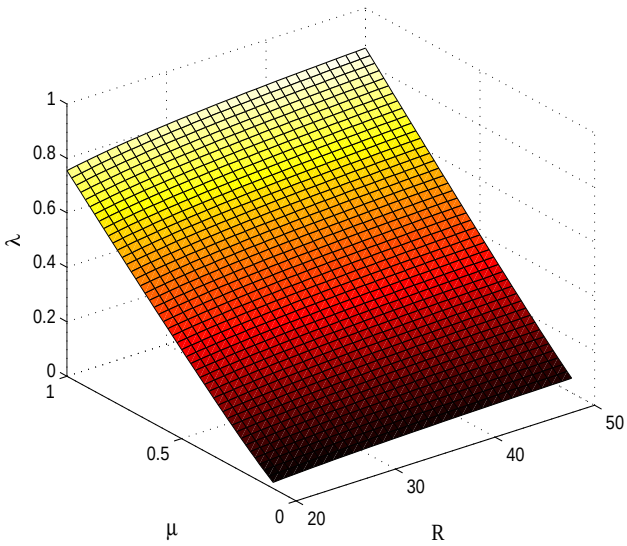

Fig. 2. Joint effect of vacation rate and service reward on λ, when $C = 1, \theta = 0.3$.

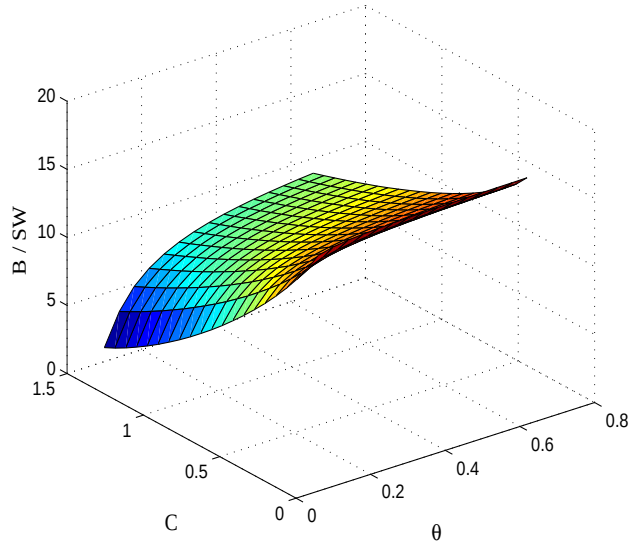

Fig. 3. Joint effect of vacation rate and per time waiting cost on B/SW, when $\mu = 1, R = 20$.

VI. Conclusion

In this paper, we discuss the optimal control of M/M/1 queueing system with setup times. Customer's equilibrium strategy and server's optimal pricing behavior are explored under two common toll structures. The first one is flat fee model and the second one is time-based fee model. Facility manager sets prices to gain the maximal profits, which is common and practical. It is found that customers in equilibrium have an identical entering rate in the two different toll models. Furthermore, the server's maximal benefit in the fixed fee model is identical with it in the time-based fee model. In addiation, the price decisions chosen by the profit-maximizing facility manager are consistent with the social to achieve social optimization. That is the objective of a profit-maximizing server and the society coincides.

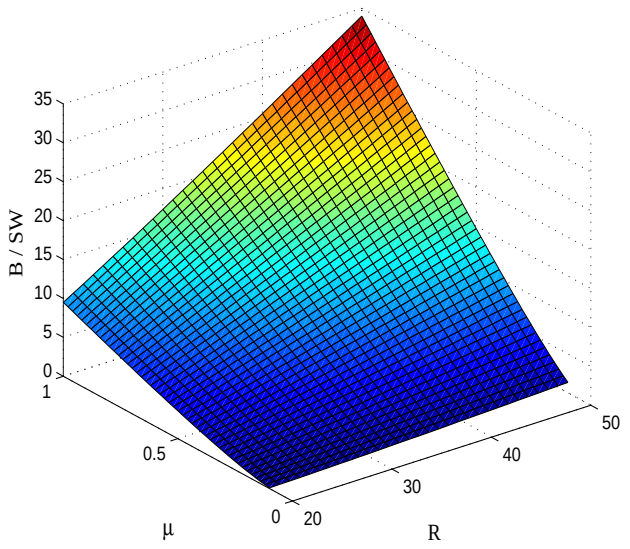

Fig. 4. Joint effect of vacation rate and service reward on B/SW, when $C = 1, \theta = 0.3$.

In all, under customers' equilibrium, the two different toll mechanisms are equivalent from the economic perspective. It is a valuable conclusion for facility managers since they can simply choose the flat fee model by charging customers with a fixed fee, instead of collecting tolls by time to avoid heavy jobs.

Those research conclusions could instruct customers to take optimal strategies and provide managers with reference information on pricing problems in queueing systems. A possible extension to this work could be to consider these pricing schemes in multiple servers queues.

REFERENCES

[1] W. Bischof, "Analysis of M/G/1 Queues with Setup Times and Vacations under Six Different Service Disciplines," *Queueing System*, vol. 39, no. 4, pp. 265-301, 2001.

[2] H. Chen, M.Z. Frank, "State Dependent Pricing with a Queue," *IIE Transactions*, vol. 33, no. 10, pp. 847-860, 2001.

[3] P. Chen, Y. Zhou, "Equilibrium Balking Strategies in the Single Server Queue with Setup Times and Breakdowns," *Operational Research*, vol. 15, no. 2, pp. 213-231, 2015.

[4] G. Choudhury, "An M^X/G/1 Queueing System with a Setup Period and a Vacation Period," *Queueing System*, vol. 36, no. 1, pp. 23-38, 2000.

[5] B.T. Doshi, "Queueing Systems with Vacations – a Survey," *Queueing System*, vol. 1, no. 1, pp. 29-66, 1986.

[6] G.V. Krishna Reddy, R. Nadarajan, R. Arumuganathan, "Analysis of a Bulk Queue with N-policy Multiple Vacations and Setup Times," *Computers & Operations Research*, vol. 25, no. 11, pp. 957-967, 1998.

[7] D.W. Low, "Optimal Dynamic Pricing Policies for an M/M/s Queue," *Operations Research*, vol. 22, no. 3, pp. 545-561, 1974.

[8] H. Mendelson, S. Whang, "Optimal Incentive-Compatible Priority Pricing for the M/M/1 Queue," *Operations Research*, vol. 38, no. 5, pp. 870-883, 1990.

[9] S. Ramasamy, O.A. Daman, S. Sani, "Discrete-time Geo/G/2 Queue under a Serial and Parallel Queue Disciplines," *IAENG International Journal of Applied Mathematics*, vol. 45, no. 4, pp. 354-363, 2015.

[10] S. Sani, O.A. Daman, "The M/G/2 Queue with Heterogeneous Servers under a Controlled Service Discipline: Stationary Performance Analysis," *IAENG International Journal of Applied Mathematics*, vol. 45, no. 1, pp. 31-40, 2015.

[11] S. Stidham, "Pricing and Capacity Decisions for a Service Facility: Stability and Multiple Local Optima," *Management Science*, vol. 38, no. 8, pp. 1121-1139, 1992.

[12] W. Sun, S. Li, "Effect of Information, Uncertainty and Parameter Variability on Profits in a Queue with Various Pricing Strategies," *International Journal of Systems Science*, vol. 45, no. 8, pp. 1781-1798, 2013.

[13] H. Takagi, *Queuing Analysis: A Foundation of Performance Evaluation in Discrete-time Systems*, North-Holland, Amsterdam, 1993.

[14] H. Takagi, *Vacation and Priority Systems*, North-Holland, Amsterdam, 1991.

[15] N. Tian, Z.G. Zhang, *Vacation Queueing Models: Theory and Applications*, Springer, Berlin, 2006.

[16] T. Wang, F. Chang, J. Ke, "Cost Analysis of a Discrete-time Queue," *Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists 2013*, IMECS 2013, 13-15 March, 2013, Hong Kong, pp. 1061-1065.

[17] M. Yajima, T. Phung-Duc, "Batch Arrival Single-server Queue with Variable Service Speed and Setup Time," *Queueing Systems: Theory and Applications*, vol. 86, no. 3, pp. 241-260, 2017.

[18] Y. Zhang, J. Wang, "Equilibrium Pricing in an M/G/1 Retrial Queue with Reserved Idle Time and Setup Time," *Applied Mathematical Modelling*, vol. 49, no. 3, pp. 514-530, 2017.