Operations Research Letters 45 (2017) 436–441

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Profit maximization in the M/M/1 queue

Refael Hassin *, Alexandra Koshman

Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel

ARTICLE INFO

Article history:
Received 7 May 2017
Received in revised form 20 June 2017
Accepted 20 June 2017
Available online 29 June 2017

Keywords:
Profit-maximization in a queueing system
Observable queues
High–low announcements

ABSTRACT

We offer a new profit-maximizing mechanism for Naor's M/M/1 model, and bound the loss incurred when the waiting room's size is limited or the server is restricted to a static price and FCFS discipline.

1. Introduction

Naor's [26] seminal work on strategic behavior in queues assumes a first-come first-served (FCFS) observable M/M/1 system with homogeneous customers. Naor shows that a fixed price is sufficient to induce socially-optimal behavior. However, a fixed price is not sufficient to maximize profits because, in general, it does not fully extract customer surplus.

This paper is about profit maximization in Naor's model. It answers the following questions:

1. *Is there a mechanism that maximizes profits and is simple to implement?* Three profit-maximizing mechanisms have been offered in the literature. They preserve the property that the queue length is observable to the customers at all times, and involve dynamic pricing or a preemptive last-come first-served (LCFS-PR) service order, both are difficult to implement. We offer a mechanism where the queue manager conceals the queue-length and only informs the customers of whether the queue length is below Naor's socially optimal threshold n^* (low congestion) or above it (high congestion). We emphasize the advantages of the new mechanism for both server and customers.
2. *How high can the loss of profit be when the queue manager is restricted to the FCFS service order and a fixed price?* We numerically study this loss and show that Naor's price may provide in the worst case only about 81% of the potential profit.
3. *How important is it to enable long queues?* We observe that a close-to-maximum profit can be obtained while strongly limiting the queue length, and bound the loss of profit when the waiting-room's size is below optimal. Maintaining a waiting room of size $k < n^*$ may cause a loss of only $k/(k+1)$ of the profit. Hence, if space is costly, a small waiting room is recommended. We also bound the gain of extending the waiting room by one unit.
4. *What if the high and low congestions are exogenously defined?* We consider high–low pricing, a restricted type of dynamic pricing. For a given threshold N, the admission fee is p_L if the number of customers in the system is at most $N-1$, and p_H otherwise. We observe that customers are encouraged to join the queue in the high-congestion states, by setting $p_H < p_L$, only if $N < n^*$. The profit loss when N is exogenously restricted is surprisingly small.

2. Literature review

Similar to our findings, Johansen [21] observes that the loss of profits associated with static pricing relative to the optimal dynamic pricing is typically small. The system considered is however different, consisting of an M/D/1 queue with observable work backlog, uniformly distributed service valuations, and an exogenous upper bound on the waiting time of admitted customers. A numerical study indicates that the loss of static control does not exceed 3%. See §2.6 in [18] for related literature.

We consider here the loss of profits when the owner of Naor's type service system is restricted in its service and pricing policies. The loss of welfare in such a system when customers' behavior cannot be controlled, often called the price-of-anarchy, is the subject of Gilboa-Freedman, Hassin, and Kerner [15]. The bounds obtained

* Corresponding author.
E-mail addresses: hassin@post.tau.ac.il (R. Hassin), alexandra.koshman@walla.com (A. Koshman).

http://dx.doi.org/10.1016/j.orl.2017.06.003
0167-6377/

there are qualitatively different from those in our case, being quite small for most values of $\rho < 1$ but infinite when $\rho > 1$.

We also raise the question of the significance of maintaining a queue and keeping waiting spaces. A similar problem is solved by Masarani and Gokturk [24] who consider an M/M/1/N queue where the server incurs a cost $C(N)$ and N is a decision variable. In their model, the customers are not delay sensitive.

Several papers consider high–low delay announcements. Allon, Bassamboo and Gurvich [3] assume that customers cannot verify the delay information provided by the firm. Altman and Jimenéz [5] assume that customers obtain high–low congestion information, but the admission fee is not changed for each case. Dobson and Pinker [11] assume heterogeneous customer waiting costs. When the system is below threshold, the queue is observable. The admission fee always stays the same. Hall, Kopalle, and Pyke [16] and Mutlu, Alanyali, Starobinski, and Turhan [25] consider threshold pricing policies, where secondary customers are blocked when the system is congested. Le Ny and Tuffin [27] consider a queueing model where a larger charge is imposed when occupancy is above threshold. Note that in our model the admission fee is *smaller* when the queue is long. Maoui, Ayhan and Foley [23] study a queue with a fixed price for queue lengths below a threshold and an infinite price above it. There are no customer waiting costs but there are server holding costs instead.

Other models assume that service rate changes when queue length exceeds a threshold. For example, Dimitrakopoulos and Burnetas [8–10] and Li and Jiang [22] discuss queueing systems in which the service rate increases when the system congestion is above threshold. Perel and Yechiali [28] consider fast and slow *phases* of service rate. During slow phases customers become impatient and may leave the queue. Chan, Yom-Tov, and Escobar [6] use a fluid model to examine conditions where speedup of service when queue length exceeds a threshold is beneficial even though faster service increases the need for rework. Shi, Shen, Wu and Cheng [29] consider a model with breakdowns. The firm changes the price between exogenous p_1 and p_2, depending on the queue length and server's state.

Economou and Kanta [12] deal with an M/M/1 model in which the waiting space of the system is partitioned into *compartments* of fixed size, and the customer is told which compartment he will enter or the position within the compartment he will have. Our system has two compartments, the first with fixed size N, the second with unlimited size, and an arriving customer is told the compartment he will enter. However, admission fees in [12] do not change for different queue lengths.

The literature on strategic models of queueing systems is surveyed by Hassin and Haviv [19] and Hassin [18]. An early version of this paper was presented in [20].

3. Mechanisms for profit maximization

Naor (1969) assumes an M/M/1 FCFS system with homogeneous risk-neutral customers arriving at rate Λ, service rate μ, service value R and waiting costs C per unit time. The maximal value of social welfare is attained if customers join the queue only when it is shorter than threshold n^*. Denote by S^* the social welfare rate under the threshold n^*. Obviously, if customers enjoy (expected) nonnegative utilities, the server's profit rate cannot be greater than S^*. A server can only attain this profit if customers join in accordance to the threshold n^* and give all their welfare to the server. We now describe three known pricing mechanisms that achieve these properties, and add a new method.

1. Chen and Frank [7] observe that a profit equal to S^* can be achieved by utilizing a FCFS regime with *dynamic pricing*, i.e., charging $p(n) = R - C\frac{n+1}{\mu}$ from a customer observing $n < n^*$ customers upon arrival, and a higher price otherwise. This pricing induces socially optimal behavior, the server receives all of the welfare generated by the system, and the net utility of each customer is equal to zero.
2. Hassin [17] shows that the LCFS-PR regime induces socially optimal customer behavior. All arriving customers join and the last customer in the queue decides whether or not to abandon the queue. Since this customer remains last until served or abandoning the queue, he imposes no externalities and his decision is socially optimal. In particular, he balks if and only if his position at the queue is $n^* + 1$. All arriving customers have the same expected utility, which is independent of queue length. Therefore, the server can obtain all the social welfare by charging the maximal price they are ready to pay.
3. A *priority pricing* mechanism for achieving S^* follows from work on priority sales by Adiri and Yechiali [1] and Alperstein [4], who showed that by adequately pricing preemptive priorities it is possible to induce threshold n^* and leave no customer surplus. An arriving customer buys the lowest priority with no current customer, and balks if all n^* priorities have customers. The result is a LCFS-PR regime, customer behavior is socially optimal, and the server's profit is S^*. An advantage of this model is that, although the outcome is LCFS-PR among customers obtaining service, customers may not feel it is unfair because they choose the type of priority to purchase. Also, those paying eventually obtain service and those balking do not incur any costs, whereas under the LCFS-PR regime with a single price the waiting costs of reneging customers are not refunded. Details can be found in Erlichman and Hassin [14].
4. We suggest a new *high–low announcements* mechanism that guarantees profit S^*. In the FCFS observable model with threshold n^*, the average customer utility for a customer arriving when $n < n^*$, is $R - CW_{<n^*}$, where $W_{<n^*}$ is the (conditional) expected waiting time of a joining customer. Therefore, $S^* = \Lambda \Pr(n < n^*)(R - CW_{<n^*})$. According to our new mechanism, the server charges a fixed price $p = R - CW_{<n^*}$ and informs the customers at any point of time whether or not $n < n^*$. Consequently, all customers join when $n < n^*$ (we assume that indifferent customers join, otherwise a slightly lower price than p will induce joining of all customers who arrive when $n < n^*$). Clearly, this guarantees the server's profit rate to be S^*.

   Note that any price greater than p can be imposed, or customers can simply be rejected, when $n > n^*$, without affecting the outcome.

Tables 1 and 2 compare the four mechanisms from the points of view of the customers and server, respectively.

Our proposed solution has advantages over the other solutions. It has a single price and avoids high switching costs, as is convenient for the server, while being fair for the customers and serves in FCFS order.

4. A high–low system

We consider an M/M/1 queueing system with a potential arrival rate Λ of risk-neutral customers, service rate μ, waiting cost rate C, and service value R. N is an exogenous constant. The queue manager sets admission fee p_L if the queue length is smaller than N (i.e., the state is L) and p_H otherwise (the state is H). An arriving customer is informed whether the state is L or H, and decides whether to join the queue or balk.

Note that both extreme cases, $N = 0$ and $N \to \infty$, lead to the unobservable version of Naor's model, as in [13].