האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

REGULATING AN OBSERVABLE M/M/1 QUEUE

By

MOSHE HAVIV and BINYAMIN OZ

Discussion Paper # 691 (September 2015)

מרכז פדרמן לחקר הרציונליות

THE FEDERMANN CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Edmond J. Safra Campus,
Jerusalem 91904, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Regulating an observable M/M/1 queue

Moshe Haviv and Binyamin Oz
Department of Statistics
and Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
91905 Jerusalem
Israel

September 20, 2015

Abstract

Naor (1969) was the first to observe that in a single-server memoryless queue, customers who inspect the queue length upon arrival and accordingly decide whether to join or not may join even if from the social point of view they are worse off. The question then is how to mechanically design the system such that customers will join only queue lengths that are advised by society, while still minding their own selfish utility. After reviewing some existing mechanisms (some involving money transfers and some not), we suggest novel ones that do not involve money transfers. They possess some advantages over the existing ones, which we itemize.

1 Introduction

Naor [9] was the first to observe that queues call for regulation: left to themselves, customers join a queue at a rate that is greater than is socially desired. The model he used is as follows. In a single-server queue there exists a Poisson arrival stream of customers at rate λ. Service times follow an exponential distribution at rate μ. All customers value service by R and suffer a cost of C per unit of time in the system (service inclusive).

The default service regime is First Come First Served (FCFS). Upon arrival, customers inspect the queue length and decide whether to join or not. They wish to maximize their mean individual monetary utility (assume without loss of generality that not joining comes with a zero utility). The customers' optimal strategy here is trivial: join if and only if the number in the system (inclusive of themselves), n, is less than or equal to n_e, where

$$n_e = \lfloor R\mu/C \rfloor. \tag{1}$$

Next Naor assumed that all costs and rewards go to a single entity (to be called "society") instead of to the individuals themselves, and that society can enforce its (socially) optimal join-do-not-join entry policy. This leads to another threshold-based policy: join if and only if the number in the system upon arrival is less than or equal to n_s for some $n_s \leq n_e$.[1] The reason behind the fact that $n_s \leq n_e$ is that when selfish customers put all costs and rewards in their equation, they ignore the negative externalities associated with joining a queue. These externalities take the form of making others (who join later) wait longer than they would have to do otherwise. Society minds these externalities, and hence the social utility associated with an individual joining customer is less than her individual utility. Therefore, joining may but need not be socially beneficial only if it is also worthy for the joiner herself; that is, the observed queue length is less than or equal to n_e.

All the above call for regulation, i.e, a set of rules, administrated by a central planner, under which the equilibrium behavior coincides with the socially optimal behavior. The purpose of this paper is to review some regulation mechanisms from the existing literature and suggest new ones.

The basic idea behind the new mechanisms suggested here is that customers should be treated differently, depending on the queue length upon their arrival. Those who arrive when the queue length is shorter than n_s are welcome by society, and should get an incentive to join, while all others are not welcome, and hence should get an incentive to balk. We would like to

[1] Naor showed that n_s is the largest integer n obeying

$$\frac{n(1-\rho) - \rho(1-\rho^n)}{(1-\rho)^2} \leq \frac{R\mu}{C},$$

where $\rho = \lambda/\mu$. Note that it is not assumed that $\rho < 1$ and in the case where $\rho = 1$ the right-hand side of the inequality above is derived by continuity and equals $(n+1)n/2$.

point out that this discrimination is not conceptually different from the situation under an unregulated system, where customers' utility is queue length dependent.

In Section 2 we review the existing and the new mechanisms. A succinct comparison of the mechanisms in this paper, with respect to some properties to be defined next, is given in Section 2.7. Finally, in Section 3, we briefly describe the unobservable version of this model as well as some regulation mechanisms that are applicable in that case.

2 Regulation schemes

The next question looked at by Naor is how to make customers adopt the socially optimal policy while still minding only their whereabouts (and without forcing them to do so).

2.1 Constant toll

Naor's suggestion is to impose a toll: each individual pays T in case of joining.[2] This payment reduces the customer's individual reward, due to receiving service, from R to $R - T$. The actual value for T (which is not unique) is selected so that the resulting (modified) n_e coincides with the existing (original) n_s. In other words, recalling equation (1), $T > 0$ is such that

$$n_s = \lfloor \frac{(R-T)\mu}{C} \rfloor.$$

One advantage of this scheme is its simplicity in the sense that the collector of the payment does not need to monitor the queue length himself. Note that T is sensitive to the model's parameters; that is, in order to determine T the organizer needs to know the value of R, C, μ, and λ.

2.2 Queue-length dependent toll

Another scheme, suggested in [11], is to charge those who observe n customers upon their arrival, when $n < n_s$, an entry fee of (almost) $R - C(n+1)/\mu$, and infinity if $n \geq n_s$. In this way individuals are left with nothing while all

[2]The entry toll does not come at any cost to society: funds collected are used elsewhere for the benefit of all.

of the consumer surplus goes to the organizer. This is also the reason why this strategy is optimal for a profit-making entity.

2.3 Purchasing priority

A scheme based on charging for priority levels that results in the same threshold of n_s was suggested in [1]. The scheme is as follows. Upon arrival a customer inspects the queue length. He has the option to balk or to purchase a preemptive priority level from a set $\{1, 2, \ldots, n_s\}$ of n_s levels. Priority level i (the lower the index, the higher the priority level) costs (a bit less than) $R - CB(i)$, where $B(i)$ is the mean busy period of an $M/M/1/i$ queue,[3] $1 \leq i \leq n_s$. Indeed, belonging to a premium class costs more. Those who pay more have a preemptive priority over those who pay less and those who belong to the same priority class are served on a FCFS basis. In [1] it is shown that the unique equilibrium strategy among customers is to purchase priority level $n_s - i$ if they see i customers upon arrival, $0 \leq i \leq n_s - 1$, and balk otherwise. In particular, the more one sees, the more one pays in case one joins. Assuming all commences from an empty system, it is possible to see that if all adopt this strategy, then whenever j customers are present in the system, $1 \leq j \leq n_s$, their purchased priority parameters are $n_s, n_s - 1, \ldots, n_s - j + 1$. An arrival who sees j customers, $0 \leq j \leq n_s - 1$, purchases priority parameter $n_s - j$ and commences service immediately (preempting the customer of priority $n_s - j + 1$ who is in service if also $j \geq 1$). Moreover, this behavior leads to a Last Come Last Served Preemption Resumed (LCFS-PR) regime being practiced. Also, the expected time in the system of a customer who purchased priority $n_s - j$ coincides with that of a busy period of an $M/M/1/n_s - j$ queue. Hence, her net utility equals zero. In particular, all the consumer surplus goes to the queue organizer as in the above-mentioned FCFS case. Society of course cares for the number of customers in the system (and not for the order in which they are served) and this is bounded by n_s, as it is socially optimal. Finally, since the customers behave in the socially optimal way and since they end up with nothing, all of the consumer surplus goes to the organizer in case he seeks profit. Moreover, since the optimal social utility bounds from above any possible profit, this is also the maximal possible profit one can make.

[3] In fact, $B(i) = \frac{1}{\mu}\Sigma_{j=0}^{i-1}\rho^j$ (see, e.g., [6], p. 137).

2.4 Not-FCFS regime

Hassin [3] suggested a regulating mechanism that does not involve money transfers. Specifically, one who joins a queue with n customers (inclusive of himself) is placed in any of the positions $1, 2, \ldots, n-1$ (but not in n). This means that she can be placed anywhere but not at the rear. It does not matter where exactly she is placed, but notice that if a joiner finds only one customer in the system (who of course is being served), the latter is being preempted while the former commences service. Interrupted service is resumed later from the point where it was last interrupted. This scheme implies that one who is at the rear of the queue stays at the rear until she is fully served, or until, and now this option is firstly mentioned, she decides to renege (abandon) due to having too many customers in front of her.[4] Of course, once she leaves the system, someone else takes her place at the rear, etc.

Leaving customers to themselves, the decision making they face now is how long should the queue ahead of them be in order from them to say "enough is enough" and decide to leave. What should this threshold be? Hassin's answer is that once customers are in position $n_s + 1$, they are better off leaving. Hassin came up with the following interesting reasoning: it is clearly socially optimal that one customer (in fact, any customer) leave at this stage, when $n_s + 1$ customers are present. More than that, since the customer at the rear does not inflict any externalities, her and society's utilities coincide. Hence, her optimal action is the same as the socially optimal one.[5]

Remark 2.1 From the point of view of regulation, it does not matter where exactly the newcomer is placed (as long as it is not at the rear), but in order to minimize the number of preemptions and to minimize the amount of lost work given to those who eventually leave without their service being completed, it is best to place the newcomer in position $n - 1$.

A clear advantage of Hassin's scheme is that it can be administrated without the administrator having to know any of the model's parameters.

[4] A customer who joins an empty system stays at the rear from her arrival until her departure (whether it is due to service completion or due to reneging). Other customers may end up at the rear due to the reneging of those who were placed behind them.

[5] For an algebraic proof see [5], p. 27–29.

2.5 Prioritizing by waiting slots

A drawback of Hassin's scheme is that it is possible that customers who have receive some service leave later without completing their service, resulting in the loss of some service effort.[6] Based on [10], we suggest the following scheme. Suppose there exist infinitely many waiting slots, numbered $1, 2, 3, \ldots$. The server always serves the one who is at the lowest indexed slot, possibly preempting the service of one who is at a higher slot. Customers, upon arrival, inspect the slots and have the option to join any vacant slot of their choice or balk for good. Once a customer occupies a slot, she cannot move to another (better) one later. From Hassin's analysis it is clear that customers join the lowest vacated slot if and only if its index is n_s or lower. Otherwise, they are better off balking. There is no loss of work in the new scheme. Preemption is still possible and can happen to anyone, with the exception of those who join slot number 1. As in [3], there is no need to know the model's parameters.

2.6 Discriminating against position $n_s + 1$

We now suggest our new regulating scheme. As before, customers inspect the queue length upon arrival. They are informed that if upon joining the number of customers present is less than or equal to n_s, they will join a high-priority class. Otherwise, they join a lower-priority class. Premium customers enjoy a preemptive priority over ordinary customers. Any regime among those belonging to the same class can be assumed. For example, it can be FCFS. It is clear now that the Nash equilibrium behavior is to join if and only if upon joining the number in the system is less than or equal to n_s. The advantages of this scheme is that there is no need for money transfers, no preemptions take place (and hence there is no loss of service effort). Moreover, the fact that a FCFS regime is applied can be looked at as a minimal switch from the current norm. The equilibrium path leads to a socially optimal admission control policy. Note, however, that it is still required to know the four parameters of the model in order to administrate this scheme. This is the case since one needs to be able to compute n_s first.

[6] Technically, this is of course not the case due to the memoryless service assumption.

2.7 Summary

All six mechanisms described above lead to socially optimal behavior. However, each mechanism comes with its own properties. In this section we compare all six mechanisms with respect to the following advantageous properties.

1. No money transfer.
2. No preemptions under equilibrium behavior.
3. No loss of work.
4. FCFS regime.
5. Robust to model's parameters.

One more criterion that we use for comparison is the ratio between the resulting customer surplus and social utility.

Scheme	Properties					$\frac{\text{Consumer surplus}}{\text{Social utility}}$
	1	2	3	4	5	
Constant toll	✗	✓	✓	✓	✗	$[0, 1]$
Queue-length dependent toll	✗	✓	✓	✓	✗	0
Purchasing priority	✗	✗	✓	✗	✗	0
Not-FCFS regime	✓	✗	✗	✗	✓	1
Prioritizing by waiting slots	✓	✗	✓	✗	✓	1
Discriminating against position $n_s + 1$	✓	✓	✓	✓	✗	1

Table 1: Summary of regulation schemes and their properties

3 Unobservable queue

Our final comment here concerns the counterpart problem when customers have to make up their mind regarding whether to join or not without inspecting the queue length first. This is Edleson and Hildebrand's [2] unobservable model. Assuming $C/\mu \leq R \leq C/(1-\rho)$, they show that customers' (Nash) equilibrium joining probability equals $p_e = (\mu - C/R)/\lambda$ while the

socially optimal joining probability is $p_s = \frac{(\mu - \sqrt{C\mu/R})}{\lambda}$. For the same reasons as those outlined in the observable version, $p_s < p_e$ (a fact that can also be checked algebraically). Imposing an entry toll of $T > 0$ such that $p_s = (\mu - C/(R-T))/\lambda$ regulates the system.[7] Other regulating schemes that are based on contracts and payments are described in [7]. It is worthwhile to mention a new scheme, suggested in [8], that involves no money transfer: give customers a random preemptive priority parameter and let them decide whether to join or not after inspecting their drawn priority level. Some related schemes are also suggested there. Finally, we would like to mention [4], which suggests a regulation scheme in which customers pay for their preemptive priority parameter.

Acknowledgement

This research was partly supported by Israel Science Foundation grant no. 1319/11.

References

[1] Alperstein, H. (1988), "Optimal pricing for service facility offering a set of priority prices," *Management Science*, **34**, 666–671.

[2] Edleson, N. M. and D. K. Hildebrand (1975), "Congestion tolls for Poisson queueing process," *Econometrica*, **43**, 81–92.

[3] Hassin, R. (1985), "On the optimality of the first come last served queues," *Econometrica*, **53**, 201–202.

[4] Hassin, R. (1995), "Decentralized regulation of a queue," *Management Science,* **41**, 163–173.

Annals of Operations Research, **113**, 15–26.

[5] Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behaviour in Queueing Systems,* Kluwer.

[6] Haviv, M. (2013), *Queueus - A Course in Queueing Theory*, Springer.

[7] Minimal algebra yields $T = R - \sqrt{CR/\mu}$.

[7] Haviv, M. (2014), "Regulating an M/G/1 queue when customers know their demand," *Performance Evaluation*, **77**, 57–71.

[8] Haviv, M. and B. Oz (2015), "Self-regulation in a queue," Manuscript submitted for publication.

[9] Naor, P. (1969), "The regulating of a queue by levying tolls," *Econometrica*, **37**, 15–24.

[10] Wang, C. L. (2015), "On socially optimal queue length," *Management Science*, Advance online publication, doi:10.1287/mnsc.2014.2148.

[11] Chen, H. and M. Frank (2001), "State dependent pricing with a queue," *IIE Transactions*, **33**, 847–860.