ON THE REGULATION OF QUEUE SIZE BY LEVYING TOLLS

by

P. Naor

Technion, Israel Institute of Technology, Haifa, Israel

and

University of North Carolina

Institute of Statistics Mimeo Series No. 493

October 1966

ABSTRACT

A queueing model -together with a cost structure- is presented, which envisages the imposition of tolls on newly-arriving customers. It is shown that frequently this is a strategy which might lead to the attainment of social optimality.

This research was supported by the Office of Naval Research under Contract No. Nonr-855(09).

DEPARTMENT OF STATISTICS

UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

On the Regulation of Queue Size by
Levying Tolls

P. Naor

Technion, Israel Institute of Technology, Haifa, Israel
and
University of North Carolina, Chapel Hill, North Carolina

I. Introduction

Some discussion has arisen recently as to whether the imposition of an "entrance fee" on arriving customers who wish to be serviced by a station - and hence join a waiting line - is a rational measure. Not much of this discussion has appeared in print; indeed this author is aware of three short communications only representing an exchange of arguments between Leeman (1964,1965) and Saaty (1965). The ideas advanced there were of qualitative character and no attempt was made to quantify the arguments.

The problem under consideration is obviously analogous to one which arises in connection with the control of vehicular traffic congestion on a road network. It has been argued* by traffic economists that the individual

---

*This author had the privilege of attending a Colloquium on "Decision Making in Traffic Planning" organized in summer 1965 by Professor Arne Jensen of the Technical University, Copenhagen. Professor Martin Beckmann of Brown University and Bonn University in his lecture at that Colloquium presented convincing arguments in favor of the thesis that the routing decision of the individual driver optimizing his own interest will not - typically - optimize an overall objective function. Hence imposing appropriate tolls may bring about optimal redistribution of vehicles moving within the road network.

car driver - on making an optimal routing choice for himself - does not optimize the system at large. The purpose of this communication is to demonstrate that indeed analogous conclusions can be drawn for queueing models if two basic conditions are satisfied:

(I) A public good is identifiable for which an objective function (typically a profit function) can be set up and maximized. This state of affairs - the existence of a public good - may manifest itself in some distinct ways two of which are: a) The population of arriving customers and the service station(s) are under the control of a single decision maker who represents the public good. b) The arriving customers represent distinct decision makers and "everybody is in business for himself." However utilities to these distinct decision makers are comparable and gains may be redistributed, e.g. through the agency of a mutual risk insurance company. Hence the expected overall profit (in unit time) accruing to arriving customers is a proper objective function representing public good.

(II) Customers are liable to be diverted from the service station, that is, some of them will be directed not to queue up (and not to invest time in that process) and not to reap the benefits available through the service at the station. This condition is in striking contrast to the usual assumptions made in queueing situations and therefore it is thought useful to elaborate a little on this point. Typically, it is assumed in most models that the mission of the station is to render service to all arriving customers (if only this does not violate the steady state condition) and an amelioration of congestion (that is: cutting of losses) may be brought about by sequencing the customers in some prescribed order. This is the rationale of most priority queueing models. In the present model the sole means of control at the disposition of the decision maker is the possible non-admission of the newly-arrived customer to the waiting line. The purpose of a toll (or an equivalent administrative measure) is precisely to prevent customers from joining the queue in case of heavy congestion and without the present condition - "customers are liable to be diverted from the service station" - there can be no rationale for the levying of tolls.

Having posed the above fundamental conditions necessary to create a framework for a queueing system with tolls we can now proceed to a detailed description of a model and a cost structure:

(i) A stationary Poisson stream of customers - with parameter λ - arrives at a single service station.

(ii) The station renders service in such a way that the service times are independently, identically, and exponentially distributed with intensity parameter μ.

(iii) On successful completion of service the customer is endowed with a reward R (expressible in monetary units). All customer rewards are equal.

(iv) The cost to a customer for staying in a queue (i.e. for queueing) is C monetary units in unit time. All customer costs are equal.

(v) The newly-arrived customer is required to choose one of two alternatives: (a) either he joins the queue, incurs the losses associated with spending some of his time in it, and finally obtains the reward; (b) or he refuses to join the queue - an action which does not bring about any gain nor loss. The decision between these two alternatives will be made by the customer through the comparison of the net gains associated with them. To avoid ambiguity it will be stipulated that in the case of a tie the customer will join the queue.

It is immediately clear that some of these assumptions represent gross simplifications of "real life" and cannot ordinarily be asserted to faithfully represent reality. Thus, for instance, there is no reason to assume that in "real life" service times are exponentially distributed, that rewards to all customers are equal (rather than statistically distributed), that queueing expenditure per unit time is identical for all customers, etc. These specific assumptions were made here since they facilitate mathematical manipulation without needlessly obscuring structure. Most important to the researcher: the pertinent feature of our model is preserved under the specific (rather than general) assumptions made, to wit, that exercise of narrow self-interest by all customers does not optimize public good.

Finally in this Section two ensuing characteristics of our model are stated:

First, it is <u>not</u> necessary to make the assumption - usual in most queueing models with a single service station - that, for steady state conditions to exist, the service intensity μ must exceed the arrival intensity λ. Arriving customers are liable to be diverted from the service station in the present model so that what is required for steady state conditions to prevail is only that the average number (per unit time) of non-diverted customers must fall short of service capacity μ. This will always be the case under the model assumptions enumerated before.

Secondly, for our model to make sense it is required that under favorable circumstances a customer will desire to queue up for (completion of) service. Hence if a newly arrived customer encounters a completely empty service station he should make the "pro-queueing" decision. His expected loss is given by $C\mu^{-1}$ whereas his reward to be collected at the end of service equals R. If the dimensionless quantity $\frac{R\mu}{C}$ is denoted by ν_s it is clear that in meaningful models the following inequality must hold

$$\nu_s = \frac{R\mu}{C} \geq 1 \qquad (1)$$

If inequality (1) does not hold the optimal policy is to disband the service station and divert the customer stream altogether.

II. Some Properties of the Model

Under the model assumptions given in the previous Section it is clear that all reasonable strategies will be of the following nature. A newly-arrived customer will observe the queue size, i say, at that instant. This quantity is a random variable whose distribution is partly determined by the strategy pursued. Now if the observed value of this random variable falls short of a constant n (the selected strategy) the newly-arrived customer will join the queue; if the observed value i is equal to n the new customer is diverted and does not join the queue. The observed value i can never exceed n in this model.

Clearly we are confronted with a system which is identical with a queueing model in which finite waiting space only is available to queueing customers. If we define

$$\frac{\lambda}{\mu} = \rho \tag{2}$$

we obtain the following steady state equations

$$p_i \rho = p_{i+1} \qquad (0 \leq i < n) \tag{3}$$

the solution of which is

$$p_i = \frac{\rho^i}{1 + \rho + \ldots + \rho^n} = \frac{\rho^i (1-\rho)}{1-\rho^{n+1}} \quad (0 \leq i \leq n) \tag{4}$$

The generating function is derived as

$$g(z) = \sum_{i=0}^{n} p_i \, z^i = \frac{1-\rho}{1-\rho^{n+1}} \cdot \frac{1-(\rho z)^{n+1}}{1-\rho z} \tag{5}$$

The expected value, q, of the random variable equals

$$q = E\,\{i\} = \frac{\rho[1-(n+1)\,\rho^n + n\rho^{n+1}]}{(1-\rho)\,(1-\rho^{n+1})} =$$

$$= \frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}} \tag{6}$$

The expected number of customers, ζ say, diverted from the service station in unit time is given by

$$\zeta = \lambda\, p_n = \frac{\lambda\, \rho^n\, (1-\rho)}{1-\rho^n} \tag{7}$$

We mention, in passing, that the busy fraction b - i.e. the degree of utilization of the service station - is, of course, not equal to ρ (as in the "usual" models) but rather

$$b = \sum_{i=1}^{n} p_i = 1 - p_o = \frac{\rho(1-\rho^n)}{1-\rho^{n+1}} \tag{8}$$

The expected number od customer joining the queue in unit time equals

$$\lambda - \zeta = \lambda(1-p_n) = \lambda \left[1 - \frac{\rho^n(1-\rho)}{1-\rho} \right] = \lambda \frac{1-\rho^n}{1-\rho^{n+1}} \qquad (9)$$

The expected number of customers leaving the service station equals

$$\mu\, b = \mu(1-p_o) = \mu \left[1 - \frac{1-\rho}{1-\rho^{n+1}} \right] \qquad (10)$$

These two quantities must be identical under steady state conditions and, indeed, it is easy to verify that

$$\rho = \frac{1-p_o}{1-p_n} \qquad (11)$$

III. Self-optimization

Let us now assume that a strategy (which will be designated as n_s) is selected in the following way (envisaged already - in a general way - in the Introduction): The newly-arrived customer weighs the two alternatives - to join or not to join the queue - by the net gains associated with them. The net gain, in the first case, is equal to

$$G_i = R - (i + 1)\, C \frac{1}{\mu} \qquad (12)$$

In the alternative case the net gain is zero. Hence self-interest is served on determining the proper value of this strategy if an integer, n_s, is found which satisfies simultaneously two inequalities

$$R - n_s C \frac{1}{\mu} \geq 0 \qquad (13)$$

and

$$R - (n_s + 1) \frac{1}{\mu} < 0 \qquad (14)$$

Inequality (13) pertains to the case where the number of queueing customers (including the one in service) encountered by the newly-arrived customer falls short of the critical number by one. The customer, of course, is supposed to join and indeed the inequality is in his favor. The inequality (14) relates to the unfavorable event: the critical number, n_s, of customers is already in the queue. We can incorporate the two inequalities in one expression

$$n_s \leq \frac{R\mu}{C} = \nu_s < n_s + 1 \qquad (15)$$

Alternatively we may express the same idea in different notation

$$n_s = [\, \nu_s \,] \qquad (16)$$

where [ ] is the well-known bracket function; that is: n_s is the largest integer not exceeding ν_o.

We note that the critical number, n_s, derived by "effectuating self-interest" depends on μ, R, and C, but not on the arrival intensity λ. This fact alone suffices - before pursuing further detailed investigation - to throw serious doubt on the social optimality of the strategy n_s.

IV. Overall Optimization

If the viewpoint is taken that the expected sum of the net gains accruing to customers in unit time is the public good which should be optimized we must proceed in a different mode from that outlined in the previous Section. We note that expected total net gain, P, under some strategy n is given by

$$P = (\lambda - \zeta)R - CE\{i\} = \lambda R(1-p_n) - Cq =$$

$$= \lambda R \frac{1-\rho^n}{1-\rho^{n+1}} - C\left[\frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\right] \qquad (17)$$

By some elementary though cumbersome considerations it can be shown that P as a function of n is "discretely unimodal" or, in other words, a local maximum is global maximum. Hence we seek that strategy, n_o say, which is associated with two inequalities

$$\lambda R\left[\frac{\rho^{n_o}(1-\rho)}{1-\rho^{n_o+1}} - \frac{\rho^{n_o+1}(1-\rho)}{1-\rho^{n_o+2}}\right] -$$

$$- C\left[\frac{(n_o+1)\rho^{n_o+1}}{1-\rho^{n_o+1}} - \frac{(n_o+2)\rho^{n_o+2}}{1-\rho^{n_o+2}}\right] < 0 \qquad (18)$$

and

$$\lambda R\left[\frac{\rho^{n_o-1}(1-\rho)}{1-\rho^{n_o}} - \frac{\rho^{n_o}(1-\rho)}{1-\rho^{n_o+1}}\right] -$$

$$- C \left[\frac{n_o \rho^{n_o}}{1-\rho^{n_o}} - \frac{(n_o+1)\, \rho^{n_o+1}}{1-\rho^{n_o+1}} \right] \geq 0 \qquad (19)$$

Lengthy but elementary considerations transform (18) and (19) into equivalent inequalities

$$R(1-\rho)^2 < \frac{C}{\mu} \left[1 - 2\rho + n_o (1-\rho) + \rho^{n_o+2} \right] =$$

$$= \frac{C}{\mu} \left[(n_o+1)(1-\rho) - \rho(1-\rho^{n_o+1}) \right] \qquad (20)$$

and

$$R(1-\rho) \geq \frac{C}{\mu} \left[n_o(1-\rho) - \rho(1-\rho^{n_o}) \right] \qquad (21)$$

These two inequalities in turn can be cast into the following form

$$\frac{n_o(1-\rho) - \rho(1-\rho^{n_o})}{(1-\rho)^2} \leq \frac{R\mu}{C} < \frac{(n_o+1)(1-\rho) - \rho(1-\rho^{n_o+1})}{(1-\rho)^2} \qquad (22)$$

To deal with (22) it will be convenient to investigate a function,

$f(\rho, \nu) = \left[\nu(1-\rho) - \rho(1-\rho^{\nu}) \right] (1-\rho)^{-2}$ of two independent variables $\rho(>0)$ and $\nu(\geq 1)$. We note, in passing that no true singularity exists for this function if $\rho=1$; rather the function is well-behaved and a non-zero and finite function value exists at that value of ρ. Next we study a situation where the value of ρ is arbitrary (positive) but fixed and we seek that unique value of ν, ν_o say, for which the function f attains the value $\nu_s (= \frac{R\mu}{C} \geq 1)$.

$$\left[\nu_o(1-\rho) - \rho(1-\rho^{\nu_o})\right](1-\rho)^{-2} = \nu_s \qquad (23)$$

It can be shown that this is always possible; furthermore since f is an increasing function of ν the integers between which ν_o lies will obey the inequalities associated with (22). Hence we arrive at

$$n_o = \left[\nu_o\right] \qquad (24)$$

an expression completely analogous to (18). Further elementary and cumbersome derivations lead to an inequality

$$\nu_o \leq \nu_s \qquad (25)$$

where the equality sign holds only * if ν_s equals unity.

V. Beneficial Toll Imposition

Inequality (25) (which typically would be strict) points to the fact that consideration of narrow self-interest deos not ordinarily lead to overall optimality. We note, of course, that even a strict inequality need not demonstrate a socially non-optimal situation since both ν_s and ν_o may be found between the same integers such that their respective bracket functions are identical. However frequently it should be expected that for the sake of

---

* The equality sign would hold also in the physically meaningless - and therefore excluded - case $\rho = 0$ (arbitrary ν_s).

narrow self-interest the facilities of the system are over-congested. To arrive at an ameliorated state of affairs it is necessary to reduce the strategy n from n_s to n_o. This can be done in two distinct ways either through an administrative rule to the effect that the maximally permissible queue size should be smaller* than a prima facie admissible number n_s; or, alternatively, a toll θ is imposed on customers joining the queue such that their (individually) expected net gain is reduced in such a way that n_o is the current criterion of newly arrived customers based on their present comparison of alternatives.

What is the optimal value, θ^*, or rather the optimal range, of the toll? Clearly this is given by

$$\frac{C}{\mu}(\nu_s - n_o - 1) = R - \frac{C(n_o+1)}{\mu} < \theta^* \leq R - \frac{Cn_o}{\mu} = \frac{C}{\mu}(\nu_s - n_o) \qquad (26)$$

If a toll taken from this range is levied on customers joining the queue the combined income (in unit time) of customers and the revenue agency is maximized. We might explicitly mention that expenditure incurred in toll collection and in information processing is considered negligible in this presentation.

Clearly, if the toll revenue may be used for redistribution of income among the population and/or for socially useful purposes the proposed imposition of tolls is an optimal procedure.

*Such a measure would have to be explained very carefully to the participants since it is in apparent contradiction with "common sense."

VI. Revenue Maximization

The toll-collecting agency may be completely divorced from the individual and collective economic interests of the customers. In that case the agency will seek to impose a toll, θ_r, designed to maximize its own revenue rather than to optimize the whole system.

The objective function of the toll-collector is given by

$$M = (\lambda - \zeta)\theta = \lambda \frac{1-\rho^n}{1-\rho^{n+1}} \left(R - \frac{Cn}{\mu}\right) =$$

$$= \lambda R \frac{1-\rho^n}{1-\rho^{n+1}} \left(1 - \frac{n}{\nu_s}\right) \qquad (27)$$

The maximization of M (which is considered a function of feasible n-s) is brought about by techniques similar to those used in previous Sections. Let the appropriate value of n be designated by n_r. It is then possible to manipulate the inequalities associated with the maximum value of M in (27) in such a fashion that a convenient quantity ν_r - analogous to ν_o in (23) - should be defined by

$$\nu_r + \frac{(1-\rho^{\nu_r-1})(1-\rho^{\nu_r+1})}{\rho^{\nu_r-1}(1-\rho)^2} = \nu_s \qquad (28)$$

The integer n_r which maximizes toll revenue is derived (as analogous integers before) by applying the bracket function on ν_r

$$n_r = [\nu_r] \qquad (29)$$

Further tedious manipulation yields

$$\nu_r < \nu_o < \nu_s \tag{30}$$

the approximative meaning of which is the following: Some toll collection may be beneficial to a queueing system if an appropriate objective function (representing public good) is chosen. However, if the toll-collecting agency is a decision maker tending to maximize its own revenue the entrance fees θ_r, levied on joining customers will be too high and social optimality will (frequently) not be attained.

$$\theta_r = R - \frac{Cn_r}{\mu} = \frac{C}{\mu}(\nu_s - n_r) \tag{31}$$

VII. Some Concluding Remarks

There is very little to add to the critique of the model and the general conclusions drawn from its structure. One point should be re-emphasized: The results -in qualitative form- are independent of the specifics of the model. Thus, for instance, if service times were distributed other than exponential we still would derive benefits from the collection of tolls, though the derivation of n_o (or an equivalent doctrine) may be much more complex than that presented in this study.

The basic features of the model are shaped by the assumption of the existence of a public good and by the assumption of possible non-admission of customers to the service station. Rewards are considered to be constant and equal. Again no basically different results would have been obtained had these rewards been drawn from a distribution. A strong modification may be called for if we were to assume that the reward obtained depends in some way on the effective traffic density. Again -without going into detailed arguments- it can be shown that a policy of "laissez faire" is only rarely and accidentally a correct one (i.e. socially optimal). In this latter more general case -in which effective interaction between customers and therefore dependence on traffic density is assumed- the proper strategy is not necessarily the imposition of a toll; cases can be constructed where the handing out of subsidies to joining customers optimizes public good. The detailed analysis of such situations is the subject of further investigation.

A series of numerical tables pertaining to the specific model presented in this study will be prepared in order to provide deeper insight into the subject matter and to facilitate the actual solution of some problems.

References

Leeman, Wayne A. — The Reduction of Queues through the Use of Price

Operations Research 12, 783-785 (1964)

Saaty, Thomas L. — The Burdens of Queuing Charges - Comment on a Letter by Leeman

Operations Research 13, 679-680 (1965)

Leeman, Wayne A. — Comments on Preceding Note

Operations Research 13, 680-681 (1965)