?/M/1: On the equilibrium distribution of customer arrivals

Amihai GLAZER
School of Social Sciences, University of California, Irvine CA 92717, U.S.A.

Refael HASSIN
Statistics Department, Tel Aviv University, Tel Aviv 69978, Israel

Received November 1981
Revised April 1982

Each day a facility commences service at time zero. All customers arriving prior to time T are served during that day. The queuing discipline is First-Come First-Served. Each day, each person in the population chooses whether or not to visit the facility that day. If he decides to visit, he arrives at an instant of time such that his expected waiting time in the queue is minimal. We investigate the arrival rate of customers in equilibrium, where each customer is fully aware of the characteristics of the system. We show that the arrival rate is constant before opening time, but that in general it is not constant between opening and closing time. For the case of exponential distribution of service time, we develop a set of equations from which the equilibrium queue size distribution and expected waiting time can be numerically computed as functions of time.

1. Introduction

Customers must wait; this is an essential feature of any interesting queuing system. But aware of this problem, each customer will behave in a manner so as to shorten the length of time he must spend waiting in queue. In particular, we shall consider a customer's decision concerning when he should visit a facility that opens at time 0, closes at time T, and uses a First-Come First-Served queuing discipline.

We are indebted to Uri Yechiali for his suggestions and comments.

North-Holland Publishing Company
European Journal of Operational Research 13 (1983) 146–150

This type of problem has been almost totally neglected in the literature; the arrival rate of customers is simply taken to be exogenously given. Such neglect is surprising because several authors have studied a related topic: how a customer's decision of whether or not to join some queue depends on the length of the queue at the time he must make his decision (see Crabill [2], Edelson [3], Knudsen [6], Naor [7], Stidham [8], and Yechiali [9]). In addition, Kleinrock [5] and Balachandran [1] studied the possibility that consumers may pay bribes to improve their position in the queue after having joined it. Finally, Jansson [4], realizing that consumers may be able to make appointments, determined the optimal appointment policy. That is, he found the instants at which customers should arrive so as to minimize the total length of time they must spend waiting.

Any customer, however, wishes to minimize only his own waiting time, and is indifferent to the effects of his acts on other customers. There is little reason, therefore, to suppose that self-interested customers would make appointments, or choose to visit the facility, at those optimal instants. Nor is there any reason to believe that customers' behavior will satisfy the modeler's assumption that the arrival rate of customers is constant over time; these problems are addressed in the following sections.

A variety of assumptions can be made about a queuing system, and presumably different models yield different results. But to clarify the central question, whether the arrival rate is constant, it is best to analyze a simple queuing model; the assumptions we use are presented in Section 2. The problem is then divided into two parts – customer arrivals prior to the facility's opening time (studied in Section 3) and arrivals from opening time to closing time (studied in Section 4). In section 5 we determine the conditions under which the arrival rate will be constant over time, and in Section 6 we give a numerical example which shows that the arrival rate need not be constant. Some conclusions are presented in the final section.

2. Assumptions

We investigate the following queuing situation. Each day the facility commences service at time zero. All customers arriving prior to time T are served during that day. A customer may obtain a favorable position in the queue by arriving before time zero, but he cannot obtain service until after that time.

The queuing discipline is strictly a First-Come First-Served one. A customer who arrives at the facility always joins the queue; balking and reneging are never worthwhile activities. The length of time required to serve a customer is exponentially distributed with mean $\bar{x}$.

Each day, each person in the population chooses whether or not to visit the facility that day. If a customer decides to visit, he also decides at what time to arrive. The total number of customer arrivals during any day is a random variable with mean n.* Each customer is fully aware of the characteristics of the queuing system and of the distributions of the variables mentioned above.

The arrival rate of customers is fully described by the function $F(t)$, namely the probability that a customer who visits the facility arrives prior to time t. A customer knows, from past experience, the expected length of the queue for any instant; no customer knows, however, the actual length of the queue until he arrives at the facility to join the queue. A customer's objective is to arrive at an instant of time such that his expected waiting time in the queue is at a minimum.

Now clearly, if all customers have this objective in mind and if we observe that one customer arrives at time t_1, and that another arrives at time t_2, then it must be that, in equilibrium, the expected waiting time is identical at these two instants. More generally, if customers arrive only during some interval, say $(-w, T)$, then this condition must hold for *any* two instants in this interval; in addition, the expected waiting time must be greater for an arrival prior to time $-w$ than for an arrival during the interval $(-w, T)$. For conciseness, we shall say that the queuing system is in equilibrium if all these conditions hold. We shall refer to the interval $(-w, T)$ as a 'day'.

Note that the expected waiting time at time t_1 depends on the pattern of customer arrivals prior to time t_1, that is, on the values of $F(t)$ for $t \leq t_1$. Thus, the requirement that the queuing system be in equilibrium imposes certain conditions on the function $F(t)$, which conditions we proceed to determine.

3. Customer arrivals before opening time

We divide our problem into two parts: the arrival of customers before time zero (at which time service begins), and after time zero. In this section we deal with the first problem, and find that the equilibrium distribution of customer arrivals is represented by the probability density function

$$f(t) = \begin{cases} 0 & \text{for } t < -w, \\ 1/n\bar{x} & \text{for } -w \leq t < 0. \end{cases} \tag{1}$$

In this expression, n is the expected number of customer arrivals during the day and w is a customer's expected waiting time. The value of n is exogenously given, but the value of w is endogenous to the model and is determined in Section 4 below.

To verify the result given in (1), consider the situation faced by a customer who arrives at time $t < 0$. The expected number of customers ahead of him in the queue is $nF(t)$, and the expected length of time required to serve them is $\bar{x}nF(t)$. But because service commences only at time zero, rather than at time t, the customer's total expected waiting time is $w = \bar{x}nF(t) - t$ (recall that $t < 0$, which explains the subtraction). In equilibrium this value must be constant for all values of t in the appropriate interval, so that a necessary condition for an equilibrium is that $d[\bar{x}nF(t) - t]/dt = 0$, or $f(t) = 1/n\bar{x}$.

Observe that service for the first customer who arrives at the facility, say at time t_1, starts precisely at time zero. Thus, such a customer's expected waiting time is simply t_1. But in equilibrium a customer's waiting time is w; we conclude that no customers will arrive prior to time $-w$. This completes our proof of (1).

For future reference, we must also determine the probability that a customer who joins the queue arrives before time zero. Integrating equation (1) we find this probability to be $\int_{-w}^{0} f(t)dt = w/n\bar{x}$.

* The derivation in Section 4 only holds when the number of arrivals during a day is a Poisson random variable.

4. Arrivals after opening time

We turn our attention to the equilibrium distribution of customer arrivals in the interval $(0, T)$. Our approach is to first determine the constraints imposed by the equilibrium conditions on the value of $f(t)$ at each instant. We then specify the transition probabilities which characterize the queuing system. These results, together with those obtained in the previous section, identify the equilibrium distribution of customer arrivals.

At this point it is useful to introduce some added notation. Let $P_k(t)$ be the probability that exactly k persons are in the system (either waiting or being served) at time t, so that $1-P_0(t)=\sum_{k=1}^{\infty}P_k(t)$ is the probability that the system is not empty at time t. Let $N(t)$ be the expected number of customers in the system at time t, and let $\mu=1/\bar{x}$ be the mean service rate.

We shall show that in equilibrium the following condition must be satisfied:

$$f(t)=\big(1-P_0(t)\big)/n\bar{x} \quad \text{for } 0\leqslant t\leqslant T. \tag{2}$$

Recall that in equilibrium a customer's expected waiting time must be the same regardless of whether he arrives at time t or at time $t+\mathrm{d}t$. But because service time is exponentially distributed, a customer's expected waiting time is simply a function of the number of persons in the queue at the time he arrives. This implies that in equilibrium $N(t)$, the expected number of customers in the system at time t, must be constant for all t, where $0\leqslant t\leqslant T$.

Now, for small $\mathrm{d}t$, $N(t+\mathrm{d}t)$ is equal to $N(t)$, plus the expected number of arrivals in the interval $(t, t+\mathrm{d}t)$, minus the probability that a customer is being served at time t and that he will leave by time $t+\mathrm{d}t$. Thus $N(t+\mathrm{d}t)=N(t)+nf(t)\mathrm{d}t-\mu(1-P_0(t))\mathrm{d}t$; the requirement that $N(t+\mathrm{d}t)=N(t)$ implies that $nf(t)=\mu(1-P_0(t))=(1-P_0(t))/\bar{x}$, which is the condition expressed in (2).

We have found the equilibrium values of $f(t)$ as a function of the values of $P_0(t)$. These latter values are endogenous to the system, and can be found by means of the transition probabilities given below:

$$P_0(t+\mathrm{d}t)=P_1(t)\,\mu\mathrm{d}t + P_0(t)(1-nf(t)\mathrm{d}t) \quad 0\leqslant t<T, \tag{3}$$

$$P_k(t+\mathrm{d}t)=P_{k-1}(t)\,nf(t)\mathrm{d}t+P_{k+1}(t)\,\mu\mathrm{d}t + P_k(t)[1-nf(t)\mathrm{d}t-\mu\mathrm{d}t] \quad k=1,2\ldots,\ 0\leqslant t<T. \tag{4}$$

To complete our system of equations, we must determine the equilibrium values of $P_k(0)$ and of w. In Section 3 we found that the probability that a customer arrives prior to time zero is $w/n\bar{x}$. Letting π_i be the probability that exactly i customers arrive during the day, we find that

$$P_k(0)=\sum_{i=k}^{\infty}\pi_i\binom{i}{k}\left(\frac{w}{n\bar{x}}\right)^k\left(1-\frac{w}{n\bar{x}}\right)^{i-k} \tag{5}$$

and

$$n=\sum_{i=0}^{\infty}i\pi_i. \tag{6}$$

Finally, the equilibrium value of w can be determined from the condition that n is the expected number of arrivals during the day. The expected number of arrivals during the interval $(0, T)$ is $\int_0^T nf(t)\mathrm{d}t$, where $f(t)$ is defined in (2). The expected number of arrivals prior to time 0 is $nw/n\bar{x}=w/\bar{x}$. Therefore, $n=w/\bar{x}+\int_0^T nf(t)\mathrm{d}t$, or

$$w=n\bar{x}-\bar{x}\int_0^T nf(t)\mathrm{d}t. \tag{7}$$

The equilibrium distribution of $f(t)$ is thus completely described by equations (1)–(7). For convenience, we repeat these equations:

$$f(t)=\begin{cases}0 & \text{for } t<-w,\\ 1/n\bar{x} & \text{for } -w\leqslant t<0,\end{cases} \tag{1}$$

$$f(t)=\big(1-P_0(t)\big)/n\bar{x} \quad \text{for } 0\leqslant t<T, \tag{2}$$

$$P_0(t+\mathrm{d}t)=P_1(t)\,\mu\mathrm{d}t + P_0(t)(1-nf(t)\mathrm{d}t) \quad \text{for } 0\leqslant t<T, \tag{3}$$

$$P_k(t+\mathrm{d}t)=P_{k-1}(t)\,nf(t)\mathrm{d}t+P_{k+1}(t)\,\mu\mathrm{d}t + P_k(t)[1-nf(t)\mathrm{d}t-\mu\mathrm{d}t] \quad \text{for } k=1,2\ldots,\ 0\leqslant t<T, \tag{4}$$

$$P_k(0)=\sum_{i=k}^{\infty}\pi_i\binom{i}{k}(w/n\bar{x})^k(1-w/n\bar{x})^{i-k} \quad \text{for } k=0,1,2\ldots, \tag{5}$$

$$n=\sum_{i=0}^{\infty}i\pi_i, \tag{6}$$

$$w = n\bar{x} - \bar{x}\int_0^T nf(t)\mathrm{d}t, \tag{7}$$

and where the parameters of the system are: $\mu = 1/\bar{x}$, the mean service rate; T, the closing time; and π_i, the probability that i customers arrive during the day.

A most important question is whether, as is usually assumed, the arrival rate is constant over time. We saw in Section 3 that during the interval $(-w, 0)$ the arrival rate is constant, but what about arrivals after the facility opens?

5. The stationary state

Under what conditions is the arrival rate stationary, that is, under what conditions is $f'(t) = 0$ for $0 \leq t < T$? To answer this question, we shall find the implications of assuming that $nf(t)$ is a constant, say c. Note from eq. (2) that $\bar{x}nf(t) = 1 - P_0(t)$, so that if $nf(t)$ is a constant, then so must be $P_0(t)$; we shall henceforth write $P_0 \equiv P_0(t)$. Imposing the requirement that $P_0(t + \mathrm{d}t) = P_0(t)$ in (3) implies that $P_1(t) = P_0 c/\mu$. We thus find that $P_1(t)$ must be a constant over time as well; moreover, employing the rule of induction on (4), we find that each variable $P_k(t)$ must also be a constant for all values of t; we can therefore write $P_k \equiv P_k(t)$ for all k. Making the successive substitutions that $P_k(t + \mathrm{d}t) = P_k(t)$ in (4), and recalling that $P_1 = P_0 c/\mu$, we find that in a stationary state

$$P_k = P_0(c/\mu)^k \quad \text{for } k = 0, 1 \ldots . \tag{8}$$

But because the P_k's are simply probabilities, we know that $\Sigma_{k=0}^{\infty} P_k = 1$, so that $\Sigma_{k=0}^{\infty} P_0(c/\mu)^k = 1$, or $P_0/(1 - c/\mu) = 1$, if $c/\mu < 1$. Making this substitution in (8) we find that

$$P_k = (1 - c/\mu)(c/\mu)^k. \tag{9}$$

We still wish to find the value of c. We make use of the fact that the expected waiting time, w, is equal to the expected number of persons in the system at time t, multiplied by the average service time per customer, or

$$w = \bar{x}\sum_{k=0}^{\infty} kP_k = (1/\mu)\sum_{k=0}^{\infty} k(1 - c/\mu)(c/\mu)^k$$

$$= \frac{c/\mu}{\mu(1 - c/\mu)}. \tag{10}$$

Note, incidentally, that this value of w is precisely the same as a customer's waiting time in a queuing system in which the facility is continuously open and in which customer arrivals can be described by a Poisson process.

Finally, by substituting (9) in (5) for the value of $k = 0$, we know that

$$1 - (c/\mu) = \sum_{i=0}^{\infty} \pi_i (1 - w/n\bar{x})^i. \tag{11}$$

If a stationary solution exists we can find it by substituting (10) in (11) and solving for c; in the next section we shall see, however, that such a solution need not exist.

6. The non-stationarity of arrivals – An example

In this section we determine the arrival rate of customers for the case in which the total number of arrivals during the day has a Poisson distribution with mean n. This is approximately the case when the population size is large and each person's decision, whether to visit the facility or not, is independent of the decisions made by any other person in the population. Recall from eq. (1) that the probability that a customer who visits the facility will do so prior to time zero is $w/n\bar{x}$, so that the number of arrivals prior to time zero has a Poisson distribution with mean $w/\bar{x}$; thus, using (5), we get

$$P_k(0) = \frac{(w/\bar{x})^k \exp(-w/\bar{x})}{k!}. \tag{12}$$

For any given value of w, we can use (2)–(4), and (12) to determine the equilibrium values of $nf(t)$ for $t > 0$. The results of such a computation are shown in Fig. 1. Finally, given these values of $nf(t)$, and given the value of w, we can always find values of n and T for which eqs. (1)–(7) are consistent with each other. That is, Fig. 1 depicts several equilibrium distributions of customer arrivals; most importantly we see that the arrival rate is not constant but rather declines over time.

This conclusion can be reached by another method; we can show in particular that $f(\mathrm{d}t) \neq f(0)$. Recall from (2) that if $f(\mathrm{d}t) = f(0)$ then it must be that $P_0(\mathrm{d}t) = P_0(0)$; making such a substitution in (3), letting $\mu = 1$, and recalling from (2) that in equilibrium $nf(0) = 1 - P_0(0)$, we find that

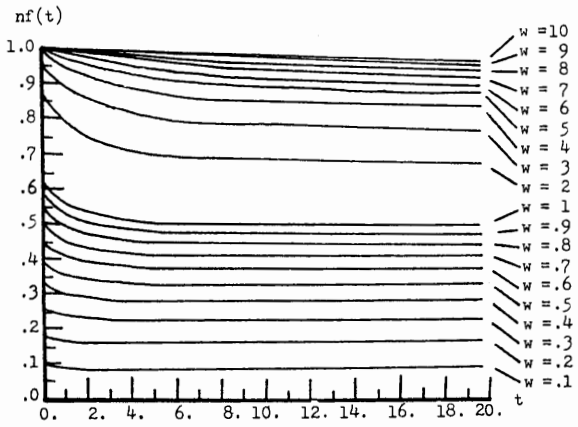

Fig. 1.

$P_0(\mathrm{d}t) = P_0(0)$ if and only if $P_1(0) = P_0(0)(1 - P_0(0))$. This condition will be satisfied if the total number of customer arrivals is geometrically distributed; indeed in this case $f(t)$ will be constant for all $t \geq 0$. It is clear, however, that if the number of customer arrivals has a Poisson distribution and if $P_0(0) > 0$, this condition is not satisfied and the arrival rate of customers is not constant over time.

7. Conclusion

We have seen that in equilibrium the expected number of customer arrivals is not constant over time. This should be of some importance in the construction of queuing models depicting facilities which are not continuously open.

But in addition, we found that individualistic behavior, in which each customer decides for himself on the time of his arrival, leads to a socially inefficient outcome. The problem is most clearly seen with respect to the behavior of customers prior to time 0. In equilibrium customers will arrive as early as w minutes prior to the opening of the facility, and an expected total of $w/\bar{x}$ customers will be wasting time standing in a queue when no one is even being served.

The problem is especially severe if n is large. Observe that in equilibrium $nf(t) = (1 - P_0(t))/\bar{x}$, which is less than $1/\bar{x}$. The expected number of arrivals during the interval $(0, T)$ is therefore

$$\int_0^T nf(t)\mathrm{d}t < \int_0^T \frac{1}{\bar{x}}\,\mathrm{d}t = \frac{T}{\bar{x}}.$$

In other words, no more than $T/\bar{x}$ customers will arrive after the facility opens, regardless of the values of n or w. Clearly, for large n relative to T and $\bar{x}$, the vast majority of customers will join the queue before the facility begins service.

This is in stark contrast to the socially optimal solution. It is clear that if the sum of customer's waiting time is to be minimized, then exactly one customer must be at the facility at time 0. But this customer will spend no time waiting, whereas on average customers must wait w minutes, so that such a situation cannot be an equilibrium one.

A solution to the problem of the non-optimality of individualistic behavior is the institution of an appointment system (see Jansson [4]). Such a solution may, however, be too expensive to institute in many situations. Another solution is service in random order. This discipline takes away all incentive to arrive before time zero. Which queuing disciplines lead to low levels of customers' waiting times? This line of research had best be left for future papers.

References

[1] K.R. Balachandran, Purchasing priorities in queues, *Management Sci. 18* (1972) 319–326.
[2] T.B. Crabill, D. Gross and M.J. Magazine, A classified bibliography of research on optimal design and control of queues, *Operations Res. 25* (1977) 219–232.
[3] N.M. Edelson and D.K. Hildebrand, Congestion tolls for Poisson queuing processes, *Econometrica 43* (1975) 81–92.
[4] B. Jansson, Choosing a good appointment system – a study of queues of the type $D/M/1$, *Operations Res. 14* (1966) 292–312.
[5] L. Kleinrock, Optimum bribing for queue position, *Operations Res. 15* (1967) 304–318.
[6] N.C. Knudsen, Individual and social optimization in a multi-server queue with a general cost-benefit structure, *Econometrica 40* (1972) 515–528.
[7] P. Naor, On the regulation of queue size by levying tolls, *Econometrica 37* (1969) 15–24.
[8] S. Stidham Jr., Socially and individually optimal control of arrivals to a $GI/M/1$ queue, *Management Sci 24* (1978) 1598–1610.
[9] U. Yechiali, Customer's optimal joining rules for the $GI/M/S$ queue, *Management Sci. 18* (1972) 434–443.