Equilibrium Arrivals in Queues with Bulk Service at Scheduled Times

A. GLAZER

University of California, Irvine, California

R. HASSIN

Tel Aviv University, Tel Aviv, Israel

In a queuing system that does not serve each customer immediately upon his arrival, a risk-neutral customer will attempt to arrive at a time that minimizes the expected length of his wait. The consequences of such behavior are explored for a queue with scheduled service.

This paper extends the conventional analysis of queues by examining a model in which each customer can choose the time at which to join a queue. We consider a system in which a finite number of customers are simultaneously served at predetermined times that follow a known schedule. Examples of such queueing systems are especially common in transportation markets. An interurban bus can carry several passengers at once, and the lengths and times of trips are usually predetermined. Airlines such as Continental operate on a schedule with planes that are often full; this may induce customers to arrive early in the hope of securing space on the next flight. Long distance ferry boat service shares these characteristics.

A customer who knows this schedule must choose an arrival time that will minimize his expected waiting costs. The length of this wait depends on the time of the customer's arrival, the capacity of the server, the frequency of service, and the behavior of all other customers. A customer can therefore be viewed as finding an optimal strategy in a noncooperative game. We show below how to find a Nash equilibrium for this game; such an equilibrium is characterized by a set of strategies such that no customer has an incentive to change his behavior. In this way we extend previous analyses: rather than treating the pattern of customer arrivals as fixed, or supposing that interarrival times are exponentially distributed, we treat arrival times as a choice variable.

A customer's expected waiting time can be divided into two parts: 1) the length of time between the instant at which he arrives and the next instant at which service is provided to someone; 2) the length of time the customer must wait because the facility is congested and cannot serve all those who happen to be in the queue. Some of our numerical results suggest that the wait until a service begins can be significant: for some reasonable values of the parameters it accounts for about 85% of the total waiting time. Instituting a random queue discipline can eliminate this wait.

A fairly extensive literature has studied customers' decisions of when to leave for an event that will start at a specified time, such as for a concert, football game, or opening of the workplace (see, for example, GAVER,[9] HENDRICKSON and KOCUR,[11] ALFA and LE MINH[1]). These papers consider the trade-offs between waiting costs, driving time, and penalties for late arrival. A time-dependent arrival pattern aimed at minimizing a customer's waiting costs was also considered by GLAZER and HASSIN[10] in their analysis of a system with known opening and closing times and a single server.

We will be concerned here with the topic of queueing systems in which customers are served in batches. Most of the literature concerning this subject is devoted to the computation of efficient control limits: a server commences service whenever a certain number of customers is in the queue. Such queues with random service time are discussed by BAILEY,[3] CHAUDHRY and TEMPLETON,[5] DEB and SERFOZO,[6] DOWNTON,[7] MEDHI,[15] NEUTS,[16] and WEISS.[20] Batch service with deterministic service times is studied by BARNETT,[4] Chaudhry and Templeton,[5] INGALL and KOLESAR,[12,13] and WEISS.[19] KOSTEN[14] lucidly describes

Transportation Science
Vol. 21, No. 4, November 1987

0041-1655/87/2104-0273 $01.25

two such queueing systems: a system in which a server with infinite capacity commences service only when a specified number of customers is in the queue, and a system in which an idle server with finite capacity commences service whenever there is at least one customer in the queue. The literature analyzing schedules is more limited; See ERLICH[8] and Weiss,[20] who, however, do not study customers' arrival decisions. TAPIERO and ZUCKERMAN[17] consider the scheduling decisions of each of two competing firms. An interesting element of their analysis lies in modeling the decisions of customers—each will go to that firm where he expects his wait to be shortest. In contrast to our analysis, however, Tapiero and Zuckerman do not consider the timing of customers' arrivals. Finally, TURNQUIST and BOWMAN[18] provide strong empirical evidence that bus passengers coordinate their arrivals with the bus schedule. They find, as our model predicts, that the distribution of arrival times is not uniform, but rather shows a peak at some time before the scheduled departure of a bus.

1. ASSUMPTIONS

CONSIDER a system in which service begins at fixed, predetermined times, regardless of the number of persons in the queue at the time a service is scheduled to begin. A bus company, for example, can schedule departures every hour on the hour, even if the bus is then empty.

We suppose that a customer's only objective is to obtain service at the lowest possible waiting cost; for present purposes we disregard any preferences customers may have for obtaining service at one time rather than another. Service starts at times $\ldots -2, -1, 0, 1, 2 \ldots$. We call each interval between two service starts a *cycle*. All customers are identical; they are served according to a first-come first-served discipline in batches that do not exceed N customers (where N is thus the server's *capacity*). Customers who happen to arrive at the same instant are assigned relative priorities by some method. If at a scheduled service time N or fewer customers are waiting in the queue, all of them will then be served. If at a scheduled service time the queue length is greater than N, exactly N customers will then be served; the rest will have to wait.

Customers arrive from a very large population. A customer first chooses the cycle during which to arrive, and then chooses his exact arrival time. He makes this decision independently of other customers and without knowing the actual length of the queue at that time; each customer, however, has full information about the probability distributions of all relevant variables (such as of the queue length and the arrival rate as a function of time). The mean number of customer arrivals in each cycle is strictly less than N.

Each customer's decision is affected by the behavior of other customers; the decision about whether to arrive, for example, at ½, ¼, or 0 time units following the most recent service depends on the distribution of the queue length at these instants. In short, we analyze a customer's arrival strategy in a noncooperative game where all players are identical and share the same information. The emphasis here will be to characterize the solution to this game, or to determine the equilibrium arrival rate function. This characterization requires the following definition.

Definition. Consider the set of all customers who arrive during some cycle. Let t be the time elapsed since the beginning of the cycle. Let $F(t)$ be the expected proportion of these customers who join the queue before time t. (Equivalently, $F(t)$ can be defined in these terms: consider some arbitrary customer who joined the queue during a cycle. $F(t)$ is the probability that he arrived before time t.) Let $w(t)$ be the expected waiting time of a customer who arrives at time $t \in (0, 1]$. Define the *support* of F as $T = \{0 < t \leqslant 1 \mid$ for all $0 < \epsilon < t$, $F(t) - F(t-\epsilon) > 0\}$. The intervals contained in T can be thought of as times during which the probability of an arrival is positive. $F(t)$ is an *equilibrium distribution* of customer arrivals if there exists a value w such that

(i) $w(t) = w$ for all $t \in T$,
(ii) $w(t) > w$ for all $t \in (0, 1]$ and $t \notin T$.

This says that if the distribution of customer arrivals is in equilibrium, the expected waiting time of all customers is identical. An example may prove instructive. Suppose buses are scheduled to depart every hour on the hour, and consider a cycle which begins immediately after 8:00 a.m. and ends at 9:00 a.m. Consider then two customers: one arrives at 8:30 and the other at 8:45. The customer who appears at 8:30 must wait half an hour until the next departure. But having arrived early, he will probably find fewer than N persons ahead of him; he is therefore likely to get space on the 9:00 bus. (It is, of course, possible that more than N persons are ahead of him, in which case he will have to wait for a bus that departs at 10:00 or even later.) The customer who arrives at 8:45 will have to wait only 15 minutes until the next bus. But his position in the queue is likely to be less favorable than that of the customer who arrived at 8:30, and he is more likely than the latter to find at least N persons ahead of him. The customer who arrives at 8:45 is therefore less likely to get a seat on the 9:00 bus, and may therefore have to wait a long time until he finds a seat.

Any customer must decide when to arrive at the station before he knows how many persons will be there at the time he arrives. A customer does, however, know the probability distribution of the number of customers in the queue at any time. Under an equilibrium distribution of customer arrivals, a customer contemplating arrival at 8:30 has the same expected waiting time as one planning to arrive at 8:45. Were the expected waiting times of customers arriving at these two instants different, at least some customers would wish to change their behavior. For example, if the expected waiting time for an arrival at 8:45 were longer than for an arrival at 8:30, some customers would choose to arrive at the earlier time. This change in behavior would change $F(t)$, the distribution of customer arrival times. Only if $F(t)$ is an equilibrium distribution does no customer have any reason to change his behavior, and thus only an equilibrium distribution can give a consistent description of customers' arrival times.

2. THE TIMING OF ARRIVALS

LET q_j be the probability that exactly j new customers arrive in a cycle of length one unit. Let r_j be the probability that exactly j persons are in the queue the instant before a scheduled service. The probability, r_0, that no persons are then in the queue is equal to the probability that the previous batch accommodated all persons in the queue (which happened if no more than N persons were in the queue) and that no one arrived during the unit time interval between scheduled departures. Thus,

$$r_0 = q_0 \sum_{i=0}^{N} r_i. \tag{1}$$

Similarly, the probability that exactly j persons are in the queue at the end of a cycle is equal to the sum of the probabilities that the following disjoint events occurred: 1) N or fewer persons were in the queue at the end of the previous cycle so that they were all served, and exactly j persons arrived during the unit length of that cycle; 2) $N + i$ persons were in the queue at the end of the previous cycle (so that N of them were served and i of them remained in the queue) and an additional $j - i$ persons arrived during the cycle. Thus

$$r_j = q_j \sum_{i=0}^{N} r_i + \sum_{i=1}^{j} q_{j-i} r_{N+i}, \qquad \text{for } j = 1, 2 \ldots. \tag{2}$$

Equations 1 and 2 define a set of simultaneous equations, whose solution can be obtained by the method of successive approximations; see BAGCHI and TEMPLETON[2] for an exposition of numerical methods, and Appendix I for a description of our solution method.

We turn now to characterize the arrival pattern within a cycle under an equilibrium distribution.

PROPOSITION 1. *Let $F(t)$ be an equilibrium distribution. Then*

(i) *$F(t)$ is continuous within each cycle.*
(ii) *Let $t_0 = \sum_{j=0}^{N-1} r_j$. Then $F(t_0) = 0$, and $F(t)$ strictly increases on $[t_0, 1]$.*
(iii) *The expected waiting time of any arrival in $[t_0, 1]$ is*

$$w \equiv \sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} r_j. \tag{3}$$

Proof. Consider part (i) of the proposition. Suppose otherwise, that $F(t)$ has a probability mass point at some point, say t_1. Consider a customer who arrives at time t_1, when with positive probability other customers also arrive. Let $p > 0$ be the probability that exactly one other customer arrives at t_1. Let $s > 0$ be the probability that the customer under consideration has lower priority than the other one who arrives at t_1. Let q be the probability that the number of persons in the system prior to time t_1 within a cycle is one less than an integral multiple of N, so that q is the probability that the wait of a customer who receives the higher priority will be 1 unit of time less than the expected wait of the other customer who arrives at time t_1.

Let $t_2 < t_1$ satisfy $(1 - t_2) < (1 - t_1) + pqs$. Then $w(t_2) < w(t_1)$, in violation of the equilibrium condition.

Part (ii) of Proposition 1 claims that the support, T, of $F(t)$ is $[t_0, 1]$. Suppose otherwise, that there exists a $t_1 \in [t_0, 1]$ and $t_2 \in (t_1, 1]$, such that $F(t)$ is constant in the interval $[t_1, t_2]$. Then the expected queue length at any time within this interval would be the same as at t_1, and a customer who arrived at t_2 would have a shorter wait than one arriving at t_1.

Since this violates the definition of $F(t)$ as an equilibrium distribution, we conclude that in equilibrium no such interval $[t_1, t_2]$ exists. It follows that $T = [t_0, 1]$ for some $0 < t_0 \leqslant 1$. The value of t_0 is determined two paragraphs below.

Since $1 \in T$, the value of w can be computed by considering a customer arriving at the end of the cycle. He will be served immediately if fewer than N persons are ahead of him in the queue; he will have to wait i units of time if upon his arrival the length of the queue is between iN and $iN + N - 1$. His expected waiting time is therefore described by Equation 3.

The waiting time of a customer arriving at t_0 consists of $(1 - t_0)$ units of time until the end of the cycle and of an additional i units of time if he finds at least iN but no more than $(i + 1)N - 1$ customers in the queue upon his arrival. Since no one arrives in $(0, t_0)$,

the length of the queue at time t_0 is between iN and $(i+1)N-1$ if the number of customers in the system at the end of the previous cycle was between $(i+1)N$ and $(i+2)N-1$. The expected wait of this customer is given by

$$w = (1-t_0) + \sum_{i=1}^{\infty} i \sum_{j=(i+1)N}^{(i+2)N-1} r_j$$
$$= (1-t_0) \sum_{i=1}^{\infty} (i-1) \sum_{j=iN}^{(i+1)N-1} r_j.$$

Making use of Equation 3 we obtain

$$1 - t_0 = \sum_{i=1}^{\infty} \sum_{j=iN}^{(i+1)N-1} r_j = \sum_{j=N}^{\infty} r_j,$$

so that

$$t_0 = \sum_{j=0}^{N-1} r_j. \qquad \square$$

To complete the analysis we must calculate the value of $F(t)$. Let $p_j(t)$ be the probability that at time t exactly j persons are in the queue; for succinctness define p_j as $p_j(t)$ for $t \in (0, t_0)$. The probability that no one is in the queue the instant after a scheduled service is the probability that N or fewer persons were in the queue just before service was scheduled; the probability that j persons remain is the probability that $N+j$ persons were in the queue. Thus,

$$p_0 = \sum_{j=0}^{N} r_j \tag{4}$$

and

$$p_j = r_{N+j}, \quad \text{for } j = 1, 2 \ldots. \tag{5}$$

In equilibrium, $dw(t)/dt = 0$ for $t_0 \leq t \leq 1$, where $w(t)$ is the expected waiting time of a customer who arrives t units of time after the beginning of a cycle. Such a customer will wait $(1-t)$ units of time until the next scheduled service, and may then have to wait further if the queue length is greater than the server's capacity. His expected waiting time is thus

$$w(t) = (1-t) + \sum_{i=1}^{\infty} \sum_{j=iN}^{iN+N-1} p_j(t), \quad \text{for } t_0 \leq t \leq 1,$$

and

$$\frac{dw(t)}{dt} = -1 + \sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} \frac{dp_j(t)}{dt}$$
$$= 0, \quad \text{for } t_0 \leq t \leq 1. \tag{6}$$

To proceed, assume that the *number* of customer arrivals during a cycle follows a Poisson distribution with mean λ. Thus arrivals follow a nonstationary Poisson process with the arrival rate $\lambda(t) = \lambda dF(t)/dt$. The probability that exactly j persons are in the queue at time $t + dt$ is then

$$p_0(t+dt) = p_0(t)[1 - \lambda(t)dt] \quad \text{for } t_0 \leq t \leq 1 \tag{7}$$

and

$$p_j(t+dt) = p_{j-1}(t)\lambda(t)dt + p_j(t)[1-\lambda(t)dt], \quad \text{for } t_0 \leq t \leq 1 \text{ and } j = 1, 2 \ldots. \tag{8}$$

From (8) we find that

$$\frac{dp_j(t)}{dt} = \lambda(t)[p_{j-1}(t) - p_j(t)], \quad \text{for } t_0 \leq t \leq 1 \text{ and } j = 1, 2 \ldots. \tag{9}$$

Substitute (9) in (6) and equate to zero to conclude that in equilibrium

$$1 = \lambda(t) \sum_{i=1}^{\infty} i[p_{iN-1}(t) - p_{(i+1)N-1}(t)]$$
$$= \lambda(t) \left[\sum_{i=1}^{\infty} i p_{iN-1}(t) - \sum_{i=2}^{\infty} (i-1) p_{iN-1}(t)\right]$$
$$= \lambda(t) \sum_{i=1}^{\infty} p_{iN-1}(t)],$$

or

$$\lambda(t) = \left[\sum_{i=1}^{\infty} p_{iN-1}(t)\right]^{-1}, \quad \text{for } t_0 \leq t < 1. \tag{10}$$

When N is equal to one, equation (10) simplifies to

$$\lambda(t) = \sum_{i=1}^{\infty} [p_{i-1}(t)]^{-1} = 1, \quad \text{for } t_0 \leq t \leq 1. \tag{11}$$

Since $\int_0^1 \lambda(t)dt = \lambda$, it follows in this case that $t_0 = 1 - \lambda$: the rate of customer arrivals is constant over the appropriate interval. For other values of N Equations 7, 8, and 10 can be solved numerically but not analytically, as described in Appendix I. Figure 1 shows some values of $dF(t)/dt$; we assume that $N = 50$ and let λ vary.

If λ is small the number of customers in the queue will rarely exceed the server's capacity, so that a waiting customer is very likely to be accommodated in the next batch. The benefit of arriving early to secure a favorable position in the queue is therefore negligible and almost all customers will arrive immediately before the start of a scheduled service. For high values of λ, however, many customers arrive early to gain a good position on the queue.

We have assumed that customers are identical, and that each one's objective is to minimize his expected waiting time. Under these conditions it is clear that the arrivals of customers before a scheduled service time are wasteful, and that customers would be better

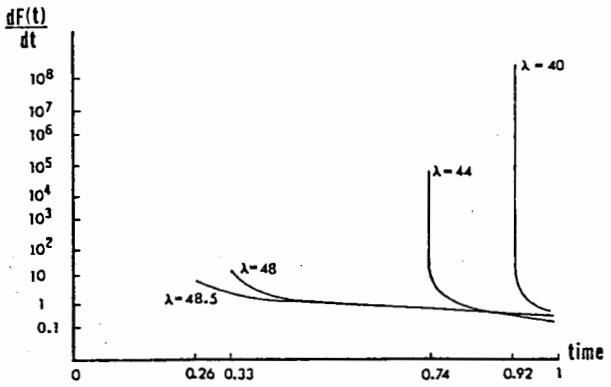

Fig. 1. Equilibrium distribution of customer arrivals ($N = 50$).

off if they all arrived only the instant before a service starts. The introduction of a random queue discipline for arrivals within a cycle, instead of a first-come first-served discipline, could lead to the desired behavior.

Column 2 of Table I shows customers' expected waiting time under scheduled service. Column 4 shows what their expected waiting would be if, in violation of the equilibrium conditions, all customers arrived only the instant before scheduled departure times. For example, when $\lambda = 46$ (and $N = 50$), we find that a typical customer's expected waiting time is 0.428. But only about 14% of this wait (0.061/0.428) is directly due to his finding the server filled to capacity; most of a typical customer's waiting time arises from his early arrival, and the subsequent wait for the next scheduled service. Although customers may view a first-come first-served discipline as fair and equitable, it causes them to waste a significant amount of time waiting in line.

APPENDIX I

THE CRITICAL part in numerically solving the equations which describe the queuing system is to determine the values of r_j. We use the method of successive approximations. Let the kth approximation, r^k, be a vector $r_1^k \ldots r_M^k$, where M is a multiple of N (we found that $M = 3N + 1$ yields sufficiently accurate results). Assign values to the initial vector, r^0, that sum to 1. For each successive approximation modify Equations 1 and 2 to obtain

$$r_0^{k+1} = q_0 \sum_{i=0}^{N} r_i^k,$$

$$r_j^{k+1} = q_j \sum_{i=0}^{N} r_i^k + \sum_{i=1}^{j} q_{j-i} r_{N+i}^k, \quad \text{for } j = 1 \ldots M.$$

The expected waiting time, w, can be calculated from these values of r substituted in Equation 3, and this value can be used for a convergence test. The time of first arrival, t_0 is computed from part (ii) of Proposition 1. The values of $p_0(0)$ and $p_j(0)$ are calculated by substituting the values of r in Equations 4 and 5; Equations 7 and 8 are used to find the values of $p_j(t + dt)$ as a function of $p_j(t)$ and $\lambda(t)$. The value of $\lambda(t)$ is found by Equation 10 using values of $p(t)$ previously determined. The average waiting time if there are no early arrivals is obtained by applying Little's formula. With probability r_{n+j} there will be j persons in the queue in a cycle. Since the average number of arrivals in a unit time interval is λ, the average wait per customer is $\sum_{j=1}^{\infty} j r_{n+j}/\lambda$. This value is shown in column (4) of Table I.

TABLE I

Waiting Times under Scheduled Service ($N = 50$)

(1) Mean No. of Arrivals λ	(2) Mean Waiting Time w	(3) Time of First Possible Arrival t_0	(4) Waiting Time if No Early Arrivals
40.0	0.078	0.922	0.005
41.0	0.108	0.892	0.008
42.0	0.147	0.853	0.012
43.0	0.197	0.803	0.018
44.0	0.259	0.741	0.026
45.0	0.336	0.644	0.039
46.0	0.428	0.572	0.061
47.0	0.539	0.462	0.098
47.7	0.634	0.372	0.147
47.8	0.649	0.359	0.156
47.9	0.665	0.345	0.166
48.0	0.680	0.332	0.177
48.1	0.699	0.317	0.190
48.5	0.761	0.267	0.241

ACKNOWLEDGMENTS

FINANCIAL SUPPORT for this research was provided by the Institute of Transportation Studies at the University of California and by the Center for Research in Organizational Efficiency at Stanford University. We are grateful to Kenneth Arrow, Kenneth Small, anonymous referees, and an Associate Editor of this journal for their comments.

REFERENCES

1. A. S. ALFA AND D. LE MINH, "A Stochastic Model for the Temporal Distribution of Traffic Demand—The Peak Hour Problem," *Trans. Sci.* **13**, 315–324 (1979).
2. T. P. BAGCHI AND J. G. C. TEMPLETON, *Numerical Methods in Markov Chains and Bulk Queues*. Springer-Verlag, Berlin, 1972.
3. N. T. J. BAILEY, "On Queuing Processes with Bulk Service," *J. Roy. Stat. Soc.* (Ser. B) **16**, 80–87 (1954).
4. A. BARNETT, "On Operating a Shuttle Service," *Networks* **3**, 305–313 (1973).
5. M. L. CHAUDHRY AND J. G. C. TEMPLETON, *A First Course in Bulk Queues*, John Wiley & Sons, New York, 1983.
6. R. K. DEB AND R. F. SERFOZO, "Optimal Control of Batch Service Queues," *Adv. Appl. Prob.* **5**, 340–361 (1973).
7. F. DOWNTON, "Waiting Time in Bulk Service Queues," *J. Roy. Stat. Soc.* (Ser. B) **17**, 256–261 (1955).
8. Z. ERLICH, "On Centralized Bus Transportation Systems with Poisson Arrivals," unpublished dissertation, School of Engineering and Applied Science, University of California, Los Angeles, 1976.
9. D. R. GAVER, JR., "Headstart Strategies for Combating Congestion," *Trans. Sci.* **2**, 172–181 (1968).
10. A. GLAZER AND R. HASSIN, "?/M/1: On the Equilibrium Distribution of Customer Arrivals," *Eur. J. Opnl. Res.* **13**, 146–150 (1983).
11. C. HENDRICKSON AND G. KOCUR, "Schedule Delay and Departure Time Decisions in a Deterministic Model," *Trans. Sci.* **15**, 62–77 (1981).

12. E. INGALL AND P. KOLESAR, "Operating Characteristics of a Simple Shuttle under Local Dispatching Rules," *Opns. Res.* **20,** 1077–1088 (1972).
13. E. INGALL AND P. KOLESAR, "Optimal Dispatching of an Infinite-Capacity Shuttle: Control at a Single Terminal," *Opns. Res.* **22,** 1008–1024 (1974).
14. L. KOSTEN, *Stochastic Theory of Service Systems*, Pergamon Press, Oxford, 1973.
15. J. MEDHI, "Waiting Time Distribution in a Poisson Queue with Bulk Service," *Opns. Res.* **16,** 189–192 (1975).
16. M. F. NEUTS, "A General Class of Bulk Queues with Poisson Input," *Ann. Math. Stat.* **38,** 759–770 (1967).
17. C. S. TAPIERO AND D. ZUCKERMAN, "Vehicle Dispatching with Competition," *Trans. Res.* **13B,** 207–216 (1979).
18. M. A. TURNQUIST AND L. A. BOWMAN, "Control of Service Reliability in Urban Bus Networks: Final Report," U.S. Department of Transportation, Report No. DOT/BSPA/DPB-50/79/5, 1979.
19. H. J. WEISS, "The Computation of Optimal Control Limits for a Queue with Batch Services," *Mgmt. Sci.* **25,** 320–328 (1979).
20. H. J. WEISS, "Further Results on an Infinite Capacity Shuttle with Control at a Single Terminal," *Opns. Res.* **29,** 1212–1217 (1981).