Schedules and Queues

UCI-ITS-WP-83-2

Amihai Glazer [1]
Refael Hassin [2]

[1] School of Social Sciences and
Institute of Transportation Studies
University of California, Irvine

[2] Department of Statistics, Tel-Aviv University

January 1983

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

ABSTRACT

In a queuing system that does not serve each customer immediately upon his arrival, a consumer will attempt to arrive at a time that will minimize the expected length of his wait. The consequences of such behavior are explored for a queue with scheduled service. The characteristics of such a system are then compared to one with bulk service, and it is found that scheduled service entails a lower expected waiting time than does bulk service except for very high values of the traffic density.

SCHEDULES AND QUEUES*

A. Glazer

R. Hassin

I. Introduction

This paper considers a class of queuing systems in which customers are served in batches, and in which the length of service is best approximated as a deterministic rather than as a stochastic variable. Such systems are common in transportation markets: an airplane can carry several passengers at once, and the length of a flight is usually predetermined; a household moving company rarely hauls the belongings of only one customer, but instead waits until it finds several shipments to the same destination. Railroads, buses, and jitney cabs possess similar characteristics.

Our major purpose is to compare customers' waiting time under two different systems. The first one involves bulk service, wherein an idle server commences service whenever a certain number of customers are in the queue. Such queues with random service times have been extensively studied by Bailey (1954), Deb and Serfozo (1973), Downton (1955), Medhi (1975), Neuts (1967), and Weiss (1979). Bulk service with deterministic service time was studied by Barnett (1973), and by Ingall and Kolesar (1972, 1974). Kosten (1973, chapter 8) lucidly describes two such queuing systems: a queue in which a server with infinite capacity

---

*Financial support for this research was provided by the Institute of Transportation Studies at the University of California, Irvine.

commences service only when a specified number of customers are waiting, and a queue in which an idle server with finite capacity commences service whenever one or more customers are waiting. We analyze a more general form of these latter models.

The other queuing system we study involves schedules: the server commences service at predetermined instants of time, regardless of the length of the queue at that time. Such a system was considered by Erlich (1976); our analysis is novel in treating a customer's decision of when to join the queue as an endogenous, rather than as an exogenous, variable.

Section II of the paper discusses bulk queues. Customer behavior in that case is simple: because customers do not know the exact times at which service will commence, each is indifferent as to when he joins the queue, and customers can be assumed to arrive according to a Poisson process.

This is not the case with scheduled service: a customer who arrives some time before service is scheduled will certainly waste some time waiting, but in return his early arrival secures him a favorable position in the queue. These two factors must be weighed, and the characteristics of the customer arrival process should be derived rather than assumed. This is done in Section III, where we find the density function of customer arrivals to be not uniform, but rather to decline over time within each service cycle. At the end of that Section we also compare customer's expected waiting times under bulk and scheduled service. We find that for all except high values of the traffic density scheduled service imposes a lower waiting cost than does bulk service.

II. Bulk Service

Consider a system in which the capacity of the single server, that is the maximum number of customers he can serve at one time, is N. The length of time required for service is one unit. Demands arrive individually according to a Poisson process with mean λ. An idle server commences service whenever n or more persons are in the queue, where $n < N$. Customers remain in the queue until they are served.

Our primary goal is to determine a customer's expected waiting time. Let q_j be the probability that exactly j new customers arrive during a unit time interval. (A summary of the notation is given in Figure 1.) The assumption of a Poisson arrival process means that

$$q_j = \lambda^j e^{-\lambda}/j! \quad . \tag{1}$$

At any instant, the server is either occupied serving customers, or else he is idle waiting to commence service again. These two alternating periods are termed a "service" period and an "idle" period. In steady state equilibrium, the probability that the queue length is zero at the end of a service period is equal to the probability of the event that no more than N persons were in the queue at the end of the previous service period (so that all customers in the queue were served once service was given) and that no new customers arrived during the service period. Let r_j be the probability that exactly j persons are in the queue at the end of a service period. Then

$$r_0 = q_0 \sum_{i=0}^{N} r_i \ . \tag{2}$$

Similarly,

$$r_j = q_j \sum_{i=0}^{N} r_i + \sum_{i=1}^{j} q_{j-i} r_{N+i}, \quad \text{for } j = 1,2,\ldots \tag{3}$$

These expressions will be used to calculate the expected waiting time of a customer.

This average wait equals the total amount of time customers spend waiting during each cycle (consisting of an idle period and of a service period), divided by the number of new customer arrivals during a cycle. Consider first customer's waiting time during a service period. With probability r_{N+j}, (for $j>0$) j persons were left behind when the service period started, and each such customer waits a unit length of time during the service period. There are an average of λ new arrivals during this period, each new arrival arriving randomly within the interval. Thus, the expected total waiting time during a service period is $\sum_{j=1}^{\infty} j r_{N+j} + \lambda/2$. The expected number of customer arrivals during such a service period of unit length is λ.

Consider next customer waiting time during an idle period. Let there be $j<n$ persons in the queue at the beginning of an idle period. Each such person waits an average $1/\lambda$ units of time until the next arrival. There will then be $j+1$ persons in the queue, and if $j+1<n$, each must wait an additional $1/\lambda$ units of time until the next arrival. Thus, total expected waiting time during an idle period, given j persons in the queue at its start, is

$$\sum_{j=0}^{n-1} r_j[j/\lambda + (j+1)/\lambda + \ldots(n-1)/\lambda] = \sum_{j=0}^{n-1} r_j(n+j-1)(n-j)/2\lambda \ . \qquad (4)$$

The expected number of arrivals during an idle period is $\sum_{j=0}^{n-1} r_j(n-j)$.

Expected waiting time, w, is the sum of these total waiting times, divided by the expected number of arrivals so that

$$w = [\lambda/2 + \sum_{j=1}^{\infty} jr_{N+j} + \sum_{j=0}^{n-1} r_j(n+j-1)(n-j)/2\lambda]/[\lambda + \sum_{j=0}^{n-1} r_j(n-j)] \ . \qquad \ldots(5)$$

Equations (1), (2), (3), and (5) define a customer's expected waiting time in the queue. Special cases of the general model are well-known, and for some values of the parameters an analytic solution can be obtained. If $n=1$, the model becomes a M/D/1 queue, for which $w=\lambda/2(1-\lambda)$. The case in which $n=N=k$ is the $E_N/D/1$ queue. If $n=1$ and $N=\infty$, we are faced with the custodian problem, discussed by Kosten (1973, pp. 119-123).

When N is large and λ is small, it is almost certain that the queue is empty at the beginning of an idle period. Service will commence once n new customers have arrived. The nth person to arrive does not wait at all, the $(n-1)$th person to arrive waits an average of $1/\lambda$ units of time, and similarly for other arriving customers. Customers' total waiting time is $(n-1)\lambda+2/\lambda+\ldots 1/\lambda = (n-1)n/2\lambda$. Since each person is as likely to be the first, second, or nth arrival, a customer's expected waiting time will be approximately $(n-1)n/2\lambda n = (n-1)/2\lambda$.

For general values of the parameters, only a numerical rather than an analytic solution can be obtained. The general features of the solution are illustrated in Figure 2. Depending on the parameters of the system, minimizing expected waiting time requires that an idle server begins service whenever at least one person is in the queue, or whenever the queue length equals the server's capacity, or at the instant the queue length has attained some intermediate value. These general results parallel those obtained for bulk service with an exponential service time (see, for example, Neuts (1967)).

If λ is relatively small, waiting time is minimized by fixing n at a small value: intuitively, it makes little sense to wait for an additional customer if he will not appear for quite some time. For large values of λ, however, waiting time is minimized by choosing large values of n. Our numerical results show, not surprisingly, that the greater the value of N, the lower the optimal value of n.

III. Scheduled Service

This section describes a different queuing system--one with scheduled service. An airline, for example, can schedule flights to depart every hour on the hour: a noon flight will leave at noon even if the plane is empty, and it will not leave at a quarter to noon even if it is by then already full.* In such a system, in contrast to one with bulk service,

---

*A different form of scheduled service arises if a plane that is filled prior to a scheduled departure time leaves immediately, instead of sitting at the terminal until that time. Under such a system, customers cannot know with certainty the time of the next departure, and this in turn will affect the timing of their arrivals. Since consumer behavior under perfect information is an extreme but interesting case, we limit our attention to scheduled service with know departure times.

customers know when a flight will depart and they can plan when to arrive at the terminal.*

A rational customer would not appear at the airport immediately after the previous flight departed; he could do better by arriving an instant before some flight. Nor will all customers arrive the instant before a flight is scheduled to depart, for with a First-Come First-Served queue discipline any one customer can dramatically improve his position in the queue by arriving an instant before a mass of other customers do. In general, then, the interarrival time cannot be characterized by an exponential distribution; we will, instead, derive the appropriate distribution. Although the model is discussed in terms of flights, it can be applied more broadly.

Let service be provided at times $-1,0,1,2\ldots$ For convenience, we speak only of one cycle, the one during the interval $(0,1)$. The server's capacity is N, and r_j is the probability that exactly j persons are in the queue the instant before a scheduled departure.

Suppose that the total number of customer arrivals during a period demarcated by two scheduled departures has a Poisson distribution; the probability of exactly j arrivals is then

$$q_j = \lambda^j e^{-\lambda}/j! \qquad (1)$$

---

*Glazer and Hassin (1982) solve a related problem: the equilibrium distribution of customer arrivals to a service facility that is open for a fixed number of hours and that adopts a FCFS queue discipline.

Note that we have made no assumptions about the distribution of customer arrivals within this unit time interval. Figure 3 depicts the relationships among the variables described.

By reasoning similar to that used in the previous section, we find that

$$r_0 = q_0 \sum_{i=0}^{N} r_i, \tag{2}$$

$$r_j = q_j \sum_{i=0}^{N} r_i + \sum_{i=1}^{j} q_{j-i} r_{N+i}, \qquad \text{for } j=1,2.... \tag{3}$$

By definition, in equilibrium all customers expect to spend the same length of time in the queue. Thus, any customer's expected waiting time, w, is the same as that of the customer arriving an instant before departure. Such a customer will find space on the next flight if there are fewer than N persons 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. The expected waiting time is therefore

$$w = \sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} r_j \, . \tag{6}$$

Equations (1), (2), (3), and (6) completely define a customer's waiting time under scheduled service, and these equations can be solved numerically; some results are presented at the end of this Section.

To complete our analysis of scheduled service, we must determine the pattern of customer arrivals, and discover how long before a scheduled departure customers arrive at the terminal. Suppose the previous departure was at time 0, and let t be the length of time that has passed since then. Let $p_j(t)$ be the probability that at time t exactly j persons are in the queue; for succintness, let p_j be defined as $p_j(0)$. The probability that no one is in the queue the instant after departure is the probability that N or fewer persons were in the queue just before departure; the probability that j persons are left behind is the probability that $N+j$ persons were in the queue. Thus,

$$p_0 = \sum_{j=0}^{N} r_j \tag{7}$$

$$p_j = r_{N+j}, \qquad \text{for } j=1,2.... \tag{8}$$

Let $f(t)$ be the probability density function of customer arrivals during the interval $(0,1)$. Define t_0 as that value of t such that $(1-t_0)$, the length of time until the next departure, plus the expected waiting time commencing at the instant a flight departs, is equal to w. Clearly in each cycle no customer will wish to arrive prior to time t_0, so that $f(t)=0$ for $0 \leq t \leq t_0$, and

$$1-t_0 + \sum_{i=1}^{\infty} \sum_{j=iN}^{iN+N-1} p_j = w \, . \tag{9}$$

Substituting (8) in (9) we find that

$$t_0 = 1-w+\sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} r_{N+j} \cdot \qquad (10)$$

The values of $f(t)$ for $t_0 \leq t \leq 1$ are found by means of expressions (11)-(15) below. Note first that the value of $f(t)$ must be positive for all values of t in this interval. Suppose otherwise, that there exist values of t_1 and dt, such that $f(t_1) > 0$, $f(t_1+dt) > 0$, but $f(t) = 0$ for $t_1 \leq t \leq t_1+dt$. A customer would clearly prefer to arrive at time t_1+dt than at time t_1, which would violate the equilibrium condition that a customer's expected waiting time is constant for all arrival times at which $f(t) > 0$.

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

$$w(t) \equiv \left[\sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} p_j(t)\right] + (1-t) \qquad \text{for } t_0 \leq t \leq 1 , \qquad (11)$$

and

$$\frac{dw(t)}{dt} = \left[\sum_{i=1}^{\infty} i \sum_{j=iN}^{iN+N-1} \frac{dp_j(t)}{dt}\right] + (1-t) \quad \text{for } t_0 \leq t \leq 1 . \qquad (12)$$

The probabilities $p_j(t)$ found in this expression, or the probability that exactly j persons are in the queue at time t is given by

$$p_0(t+dt) = p_0(t)[1-\lambda f(t)dt] \qquad \text{for } t_0 \leq t \leq 1 \tag{13}$$

and

$$p_j(t+dt) = p_{j-1}(t)\lambda f(t)dt + p_j(t)[1-\lambda f(t)dt] ,$$
$$\text{for } t_0 \leq t \leq 1 \text{ and } j=1,2\ldots, \tag{14}$$

from which we find that

$$\frac{dp_j(t)}{dt} = \lambda f(t)[p_{j-1}(t) - p_j(t)], \qquad \text{for } t_0 \leq t \leq 1. \tag{15}$$

Substituting (15) in (12), we conclude that in equilibrium

$$f(t) = \left[\lambda \sum_{i=1}^{\infty} p_{iN-1}(t)\right]^{-1} \qquad \text{for } t_0 \leq t \leq 1 . \tag{16}$$

For $N=1$ these equations simplify to $f(t) = \left[\lambda \sum_{i=1}^{\infty} p_{i-1}(t)\right]^{-1} = 1/\lambda$, and $t_0 = 1-\lambda$: the rate of customer arrivals is constant over the appropriate interval. For other values of N these equations can be solved numerically but not analytically. Some values of $f(t)$ are shown in Figure 5.

If λ/N 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 by the next flight. The benefit of arriving early to secure a favorable position in the queue is therefore negligible and almost all customers will arrive immediately before a scheduled departure. For high values of λ/N, however, many customers arrive early to gain a good position on the queue.

This early arrival is in some sense a waste; everyone would be better off if all customers arrived only the instant before a departure, and no earlier. The introduction of a random queue discipline, instead of a FCFS one, could lead to that result. Nor is this wasted waiting time insignificant. Figure 6 shows the fraction of total waiting time attributable to customers early arrivals, or the value of $\int_0^{t_0} f(t)dt/w$. For some reasonable values of the parameters, this fraction will be greater than one-half. 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.

Our final objective is to find which of the two queuing systems discussed, bulk service or scheduled service, imposes a lower waiting time on customers. Figure 4 shows expected waiting time under the two systems for various values of λ and N.* Note that bulk service is always superior to scheduled service for $N=1$. Such is not the case if $N \geq 2$. Instead, there exists a function, $\lambda^*(N)$, such that if

---

*The expected waiting times under bulk service are calculated under the assumption that the optimal value of n is chosen for each λ and N.

$\lambda > \lambda^*(N)$ bulk service entails a lower waiting time than does scheduled service; if $\lambda < \lambda^*(N)$ scheduled service is superior to bulk service. Figure 7 depicts these critical values of λ for various values of N. Note that the ratio $\lambda^*(N)/N$ is quite high; if $N=6$, for example, $\lambda^*(N) \approx 5.6$, and bulk service is better than scheduled service only if $\lambda/N > 0.93$. That is to say that scheduled service is the preferred queuing system unless the traffic density, and the consequent congestion, is quite high.

References

Bailey, N.T.J. (1954), "On Queuing Processes with Bulk Service," Journal of the Royal Statistical Society, Series B, vol. 16, pp. 80-87.

Barnett, A. (1973). "On Operating a Shuttle Service." Networks, vol. 3, no. 4, pp. 305-313.

Deb, R.K. and R.F. Serfozo, (1973), "Optimal Control of Batch Service Queues," Advances in Applied Probability, vol. 5, pp. 340-361.

Downton, F. (1955), "Waiting Time in Bulk Service Queues," Journal of the Royal Statitistical Society, Series B, vol. 17, pp. 256-261.

Erlich, Zipora (1976), "On Centralized Bus Transportation Systems with Poisson Arrivals," unpublished dissertation, School of Engineering and Applied Science, University of California, Los Angeles.

Glazer, A. and R. Hassin (1982), "?|M|1: On the Equilibrium Distribution of Customer Arrivals," forthcoming, European Journal of Operational Research.

Ingall, E. and P. Kolesar (1972), "Operating Characteristics of a Simple Shuttle under Local Dispatching Rules," Operations Research, vol. 20, no. 6, pp. 1077-1088.

Ingall, E. and P. Kolesar (1974), "Optimal Dispatching of an Infinite-Capacity Shuttle: Control at a Single Terminal." Operations Research, vol. 22, no. 5, pp. 1008-1024.

Kosten, L. (1973), Stochastic Theory of Service Systems. Oxford: Pergamon Press.

Medhi, J. (1975), "Waiting Time Distribution in a Poisson Queue with Bulk Service," Operations Research, vol. 16, pp. 189-192.

Neuts, M.F. (1967), "A General Class of Bulk Queues with Poisson Input," Annals of Mathematical Statistics, vol. 38, pp. 759-770.

Weiss, H. J. (1979) "The Computation of Optimal Control Limits For a Queue with Batch Services," Management Science, vol. 25, no. 4, pp. 320-328.

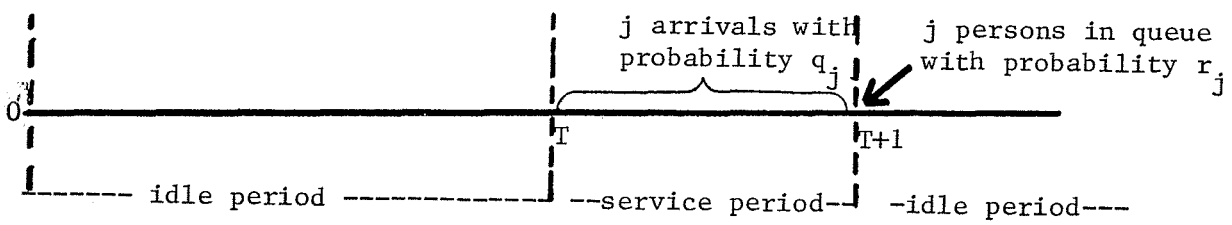


Figure 1

Notation for Bulk Service Queue

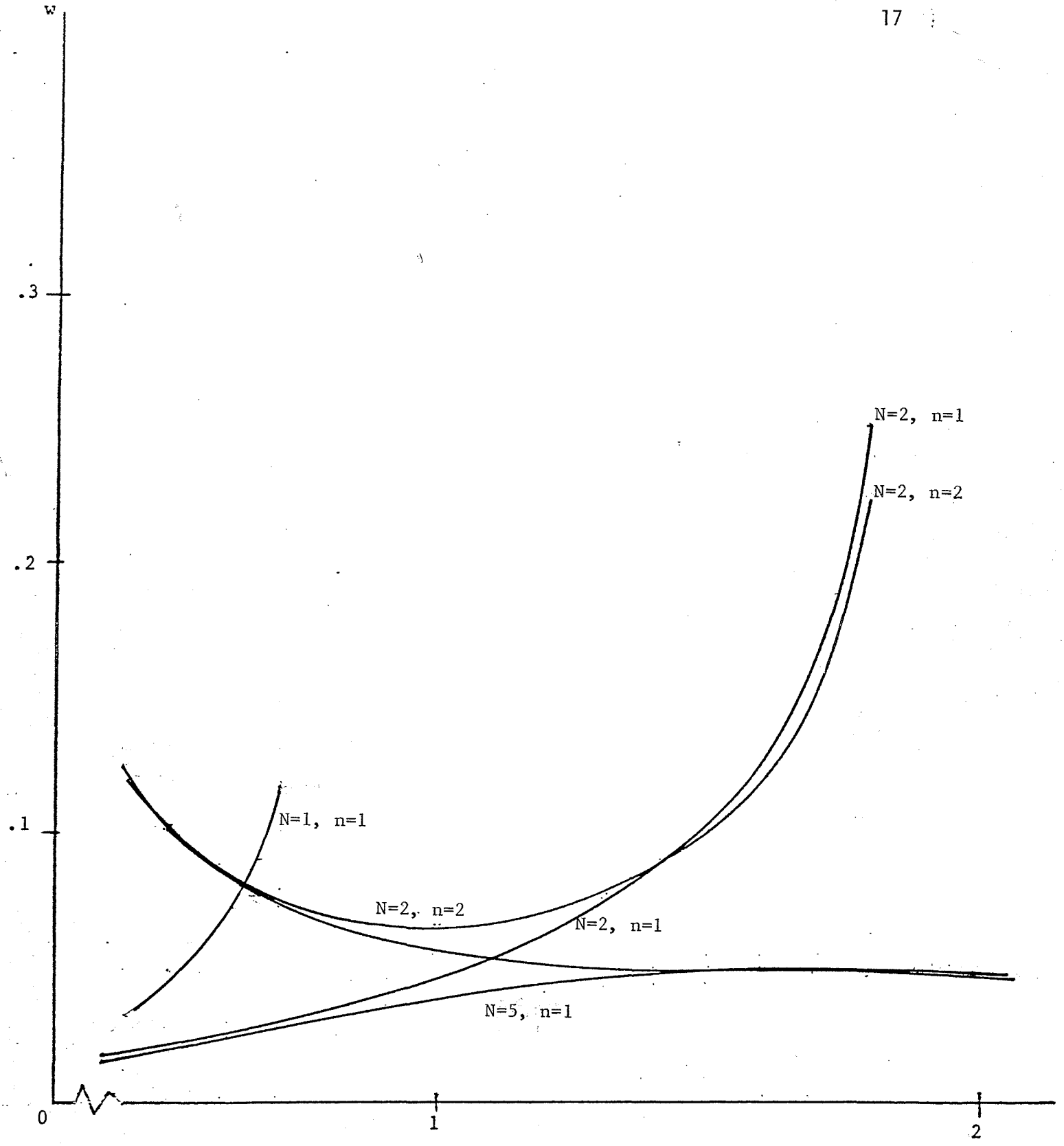


Figure 2

Waiting time with bulk service for various values of n, N, and λ.

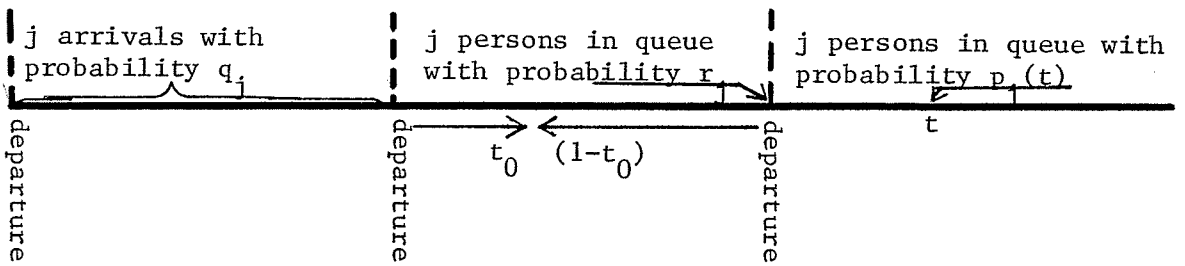


Figure 3

Notation for Scheduled Service Queue

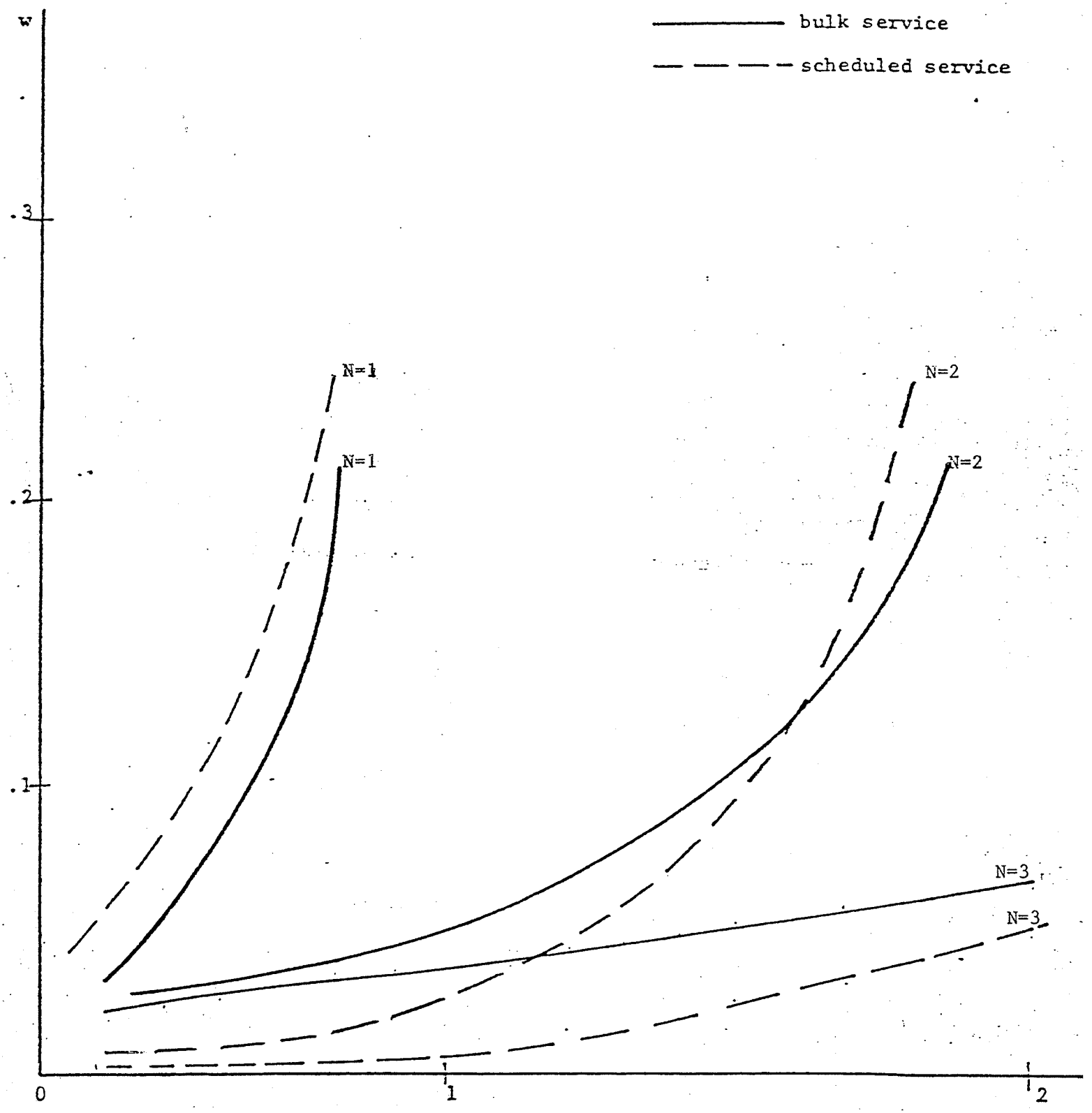
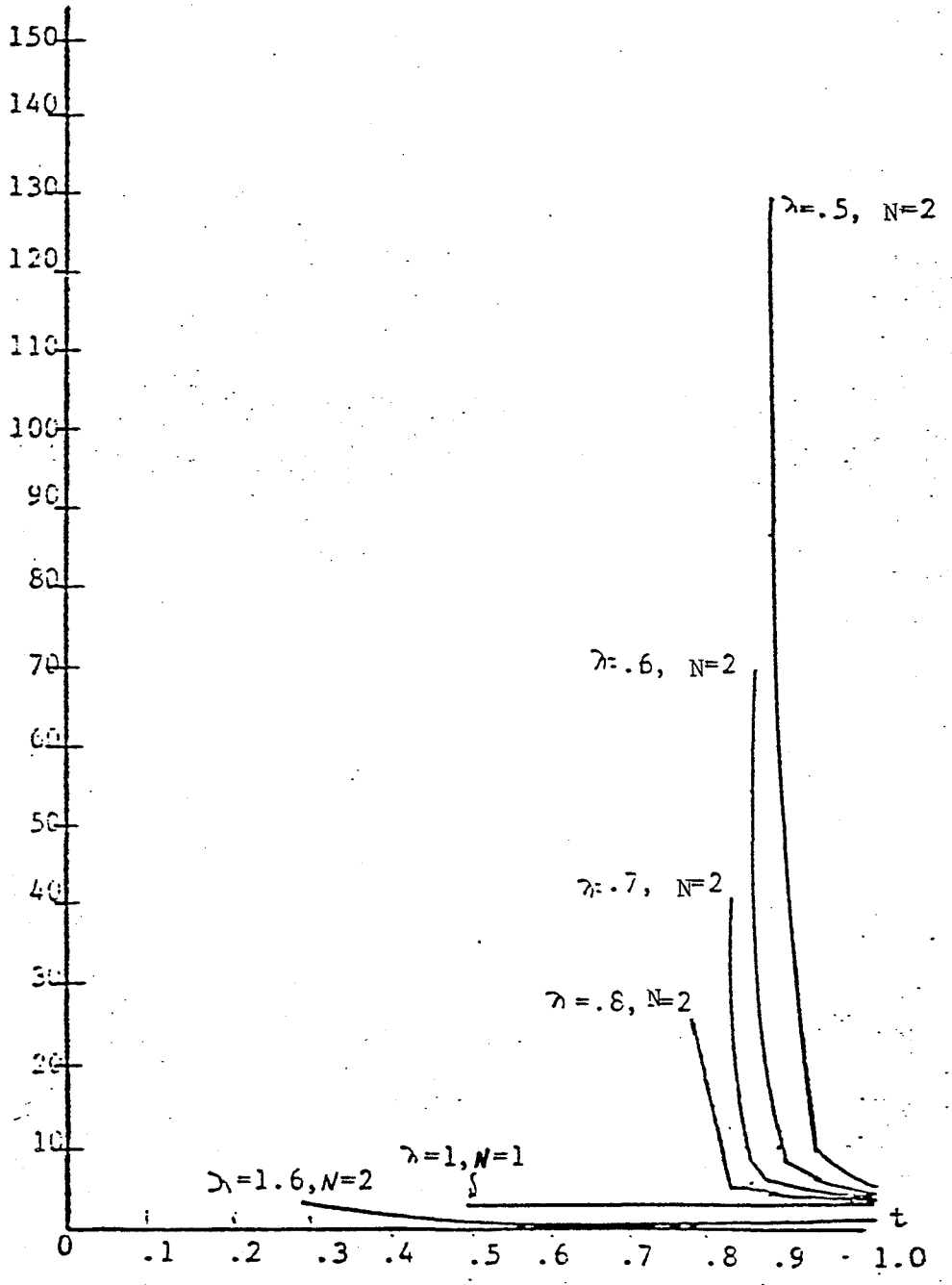


Figure 4

Expected waiting times



Distribution of Customer arrivals under scheduled service

Figure 5

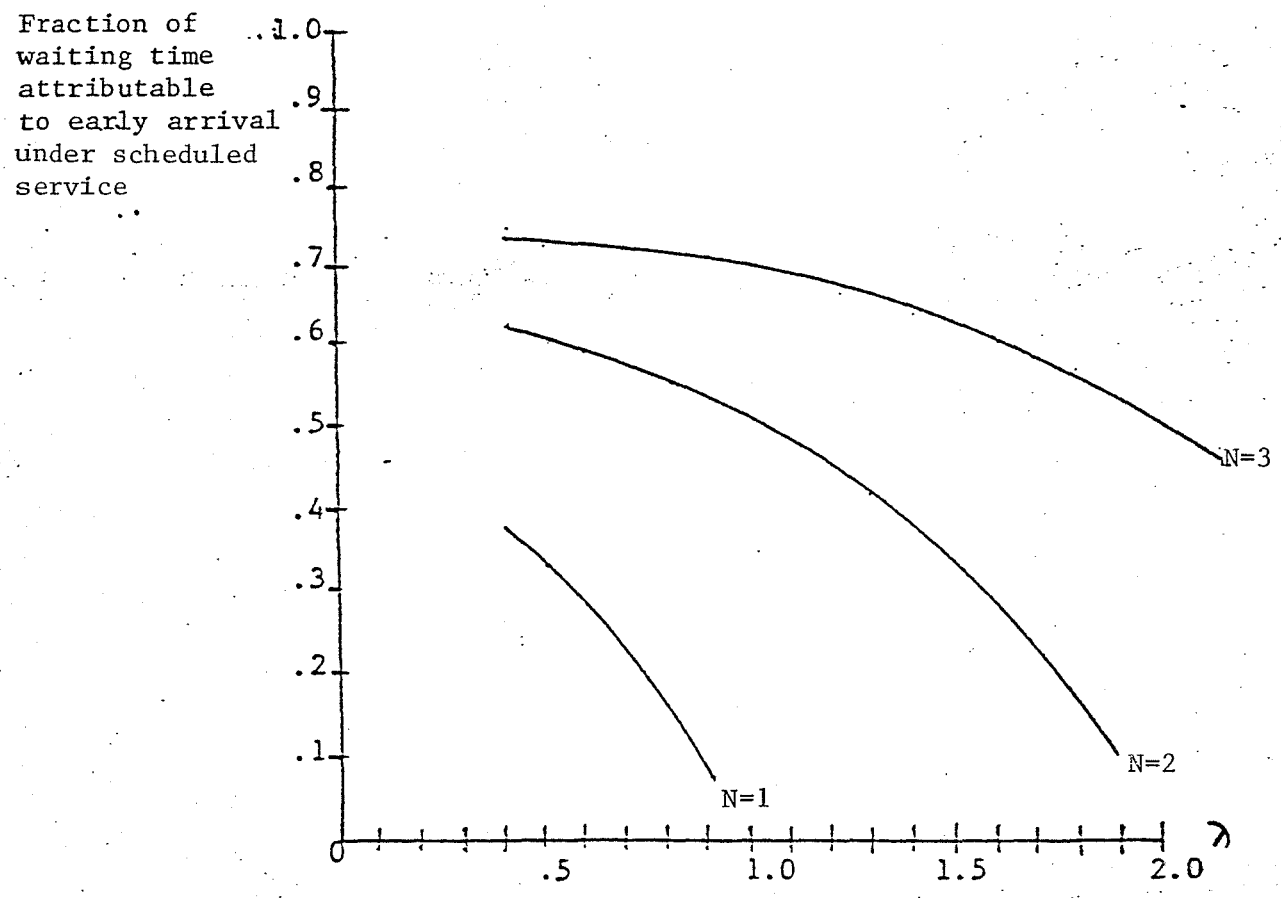

Fraction of waiting time attributable to early arrival under scheduled service
1.0
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
.5
1.0
1.5
2.0
λ
N=3
N=2
N=1

Figure 6

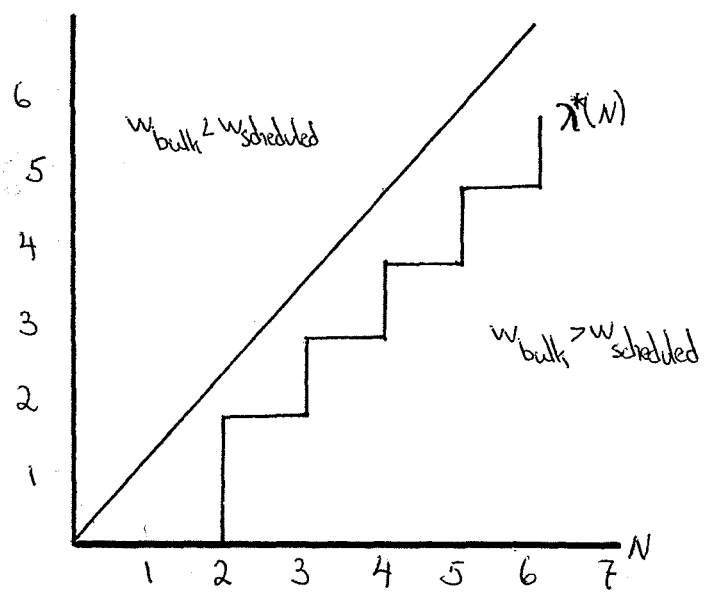

6
5
4
3
2
1
$w_{built} < w_{scheduled}$
$\lambda^*(N)$
$w_{built} > w_{scheduled}$
1
2
3
4
5
6
7
N

Figure 7