Static and Dynamic Scheduling of Customer Arrivals to a Single-Server System

P. Patrick Wang
Mathematics Department, The University of Alabama, Tuscaloosa, Alabama 35487

In this article we consider a single-server system whose customers arrive by appointments only. Both static and dynamic scheduling problems are studied. In static scheduling problems, one considers scheduling a finite number of customer arrivals, assuming there is no scheduled customer arrival to the system. In dynamic scheduling problems, one considers scheduling one customer arrival only, assuming that there are a number of scheduled customers already. The expected delay time is recursively computed in terms of customer interarrival times for both cases. The objective is to minimize the weighted customer delay time and the server completion time. The problem is formulated as a set of nonlinear equations. Various numerical examples are illustrated. © 1993 John Wiley & Sons, Inc.

1. INTRODUCTION

In many queueing systems, customers arrive to the system for service not randomly, but rather in a manner determined ahead of time. For instance, patients need to make appointments before they see the doctor. In a manufacturing system, one may want to load parts onto the shop floor at different times in order to reduce the working-in-process inventory and minimize the system completion time. Since the service times required by customers are unknown, the interarrival times are usually set with equal time intervals. It is natural then to ask: Does this equally spaced interarrival time give the best system performance? If not, what are the best times for customers to enter the system? Due to randomness of the service process, we cannot control the service process, but exact customer arrival times can be controlled. The problem is then how to arrange customer arrivals such that certain objectives are optimized. This article attempts to answer this question.

Although a great deal of attention has been paid to queueing control models, little has been done concerning the above problem. The scheduling problem of a $D/M/1$-type queueing model was studied by Jansson [5], in which the author studied the optimal constant interarrival time. In Mercer [7, 8], the limiting probabilities of the queue length and waiting were analyzed when the customers are scheduled to arrive, but the actual arrival times are allowed to differ from the schedule. In [11], Sabria and Daganzo examined the same problem as in [7] and developed simple expressions in terms of a few parameters for the waiting

Naval Research Logistics, Vol. 40, pp. 345–360 (1993)
 CCC 0894-069X/93/030345-16

time distribution. This model has important applications in ship competition for berth space at a seaport.

All of the articles above assume an infinite number of customers and constant interarrival times. The research focus has been on the steady-state probabilities. In reality, systems rarely reach a steady state, as is the case where a doctor's office is operating from a new initial state each working day. The scheduling of a finite number of arrivals is considered by Rosenshine [10] and Healy [4]. Rosenshine named this model as $S(n)/M/1$, where $S(n)$ represents scheduling n arrivals, the letter M represents exponential service times, and there is only one server. The author also proved that the expected total waiting time is a convex function of the customer interarrival times when the number of customers is no more than three. In [4], the author solved the same problem by using computer simulation. Such a method allows any kind of service distributions. The perturbation analysis was introduced to generate the gradient during each run of the simulation.

In this article we consider the transient solution of an $S(n)/M/1$ type of queue and attempt to determine the exact arrival times for a finite number of customers. The basic system performance considered here is a combination of customer delay time (waiting time plus service time) and the server completion time. Two problems are investigated. One is a "static" scheduling problem in which a finite number of customers are scheduled at one time. It is static in the sense that the schedule is obtained once. The other is a "dynamic" scheduling problem in which only one customer is scheduled given that there are a certain number of already scheduled customers in the system. Consider the situation where the system has been under operation for a while and a new customer calls for an appointment; what is the best arrival time such that the objective function is optimized? We assume this new customer that needs to be scheduled is independent of the scheduled customers and can be "inserted" anywhere between customers. The dynamic scheduling problem studies the optimal arrival time for this new customer.

This article is organized as follows. Section 2 gives the system assumptions and the problem formulation. It briefly reviews the phase-type (PH) probability distribution and illustrates that the customer delay time distribution can be represented by a PH distribution. Section 3 deals with the static scheduling problem in which the expected customer delay time and the derivative of the delay time with respect to the customer interarrival times are derived. The optimal interarrival times are determined by solving a set of nonlinear equations. Section 4 solves the dynamic scheduling problem. Numerical examples are given in Section 5. General service time distributions and further extensions of the research are discussed in Section 6.

2. ASSUMPTIONS AND NOTATION

Assume that there is a single server who works whenever work is available and is idle otherwise. The service times are independent and identically distributed (iid) random variables with an exponential distribution. Without losing generality, assume the server works at unit rate. Customers arrive to the server by appointment and once the appointment is made, the customer will arrive

punctually. Rescheduling is not permitted. The queueing discipline is first-come-first-serve (FCFS). At time zero, there are n customers that need to be scheduled and there are k (>0) customers in the system waiting to be served. For the case the system is empty at time zero ($k = 0$), we can always let one customer come at time zero and then schedule the remaining ($n - 1$) customers.

Denote $t_0 = 0$ and t_i, $i = 1, \ldots, n$, customer i arrival time. Let x_i, $i = 1, \ldots, n$ be the interarrival time between customer $(i - 1)$ and customer i; i.e.,

$$x_i = t_i - t_{i-1}, \qquad i = 1, 2, \ldots, n.$$

Let W_i be the random delay time (waiting time plus service time) of customer i. Let w_i be the expected delay time of customer i; i.e., $w_i = EW_i$. Let $\bar{x} = [x_1\ x_2 \cdots x_n]^T$ and $\bar{w} = [w_1\ w_2 \cdots w_n]^T$ be the interarrival time vector and the expected delay time vector, respectively. Finally, let $w = \Sigma_{i=1}^{n} w_i$ and $N = n + k$.

The general form of the problem is

$$\begin{aligned} &\underset{\bar{x}}{\text{minimize}}\ J(\bar{x}, \bar{w}), \\ &\text{subject to } \bar{x} \in \Omega, \end{aligned} \tag{1}$$

where $\Omega \in R^n$ is a feasible set of $\bar{x}$. If the objective function $J(\bar{x}, \bar{w})$ is a linear combination of the expected total customer delay time and the service completion time, then we obtain the static problem,

$$\begin{aligned} \text{(STATIC)} \qquad &\text{minimize } J(\bar{x}, \bar{w}) = \alpha u^T \bar{w} + (1 - \alpha)(u^T \bar{x} + u_n^T \bar{w}), \\ &\text{subject to } \bar{x} \geq 0 \end{aligned} \tag{2}$$

where $0 \leq \alpha \leq 1$ is the weighting factor, u is a column vector with all elements 1, and u_n is a vector with its nth element set to one and all others zero. The first term in (2) is the total customer delay time, and the second term is the system completion time which is the completion time of customer n, i.e., $t_n + w_n$.

For special values of the weighting factor α, the solution $\bar{x}$ is obvious. For example, if $\alpha = 0$, then $\bar{x} = \bar{0}$, which means we are trying to complete the n services as soon as possible regardless of the customers' delay time. All of the n customers are scheduled to come to the system at time 0. If, on the other hand, $\alpha = 1$, then $\bar{x} = \infty$, which means that the customer waiting time is much more important than the server completion time. Customers are scheduled to come sufficiently separated so that no customer will wait.

Next, we give the definition of the phase-type (PH) distribution. Detailed discussion of PH distribution can be found in Neuts [9]. A PH distribution $F(x)$ is the distribution of the absorption time in a finite Markov process defined by the matrix

$$\begin{bmatrix} T & T^0 \\ 0 & 0 \end{bmatrix},$$

where the square matrix T satisfies $T_{ii} < 0$, and $T_{ij} \geq 0$, for $i \neq j$, and the column vector T^0 satisfies $Tu + T^0 = 0$. The initial probabilities of the Markov process are given by a row vector a.

The pair (a, T) is called a representation of $F(x)$. The mean of $F(x)$ is readily computed by

$$EX = aT^{-1}u. \tag{3}$$

A large variety of distributions can be represented by the PH distribution [9]. One of the major advantages of PH distributions is structural and computational. Instead of dealing with differential equations, complex variables, and numerical integrations, PH distributions can be handled using the matrix method. For instance, the generalized Erlang distribution of order m with parameters λ_1, $\lambda_2, \ldots, \lambda_m$ has the representation (a, T), where $a = [1, 0, \ldots, 0]$ and

$$T = \begin{bmatrix} -\lambda_1 & \lambda_1 & & & & \\ & -\lambda_2 & \lambda_2 & & & \\ & & & \cdots & & \\ & & & & -\lambda_{m-1} & \lambda_m \\ & & & & & -\lambda_m \end{bmatrix}.$$

It should be pointed out that other efficient algorithms for computing the steady-state probabilities of various queueing models are available; for example, see Chaudhry, Harris, and Marchal [1] and Grassmann and Chaudhry [2]. We use PH distribution here due to its matrix structure.

3. STATIC SCHEDULING PROBLEM

In a static scheduling problem, we assume that no customers have been scheduled to arrive to the system. The scheduling list is empty. The task of this section is to reveal the structure of the expected delay times, $w_i(\bar{x})$, in terms of the interarrival times $\bar{x}$. We first construct an embedded Markov chain and then show that the random variable W_i has a PH distribution.

The embedded Markov chain $\{C_i, i \geq 0\}$ is defined as the number of customers in the system (including the one in service) right before the epoch of a customer arrival. An equivalent definition of the embedded Markov chain is that C_i is the number of customers in the system observed by customer i at the arrival epoch. When customer i comes to the system and observes m customers in the system, then the delay time of customer i, W_i, is a sum of $(m + 1)$ iid service times which has an Erlang distribution with phase $(m + 1)$. Since there are k customers at time zero and only a finite number n of customer arrivals are considered, the number of customers observed at the epoch of the arrival of customer i can not exceed $(k + i - 1)$ for any i. Therefore, the possible states of the embedded Markov chain are $0, 1, 2, \ldots, k + n - 1$ and the total number of states is N (recall $N = k + n$). We only need to consider the first n state transitions.

Let $p_i(m) = \Pr\{C_i = m\}$, for $0 \leq m \leq N - 1$ and $1 \leq i \leq n$, be the probability that customer i observes m in system; then the probability distribution function

of W_i is just a mixture of Erlang distributions which can be represented by a PH representation $(\bar{p}_i, S)$, where

$$S = \begin{bmatrix} -1 & & & & \\ 1 & -1 & & & \\ & & \cdots & & \\ & & & -1 & \\ & & & 1 & -1 \end{bmatrix} \tag{4}$$

is an $N \times N$ matrix, and

$$\bar{p}_i = [p_i(0), p_i(1), \ldots, p_i(N-1)] \tag{5}$$

is a row vector of order N. The expected delay time is then easily obtained from Eq. (3):

$$w_i = -\bar{p}_i S^{-1} u. \tag{6}$$

Noting $S^{-1}u = -[1, 2, 3, \ldots, N]^T$ and letting $\beta = -S^{-1}u$, we obtain w_i as a dot product of $\bar{p}_i$ and β,

$$w_i = \bar{p}_i \beta, \qquad i = 1, 2, \ldots, n. \tag{7}$$

Or a summation form is

$$w_i = \sum_{m=0}^{N-1} p_i(m)(m+1). \tag{8}$$

Next, we show that $\bar{p}_i$ can be computed recursively on $\bar{p}_{i-1}$. If m customers are in the system at observation time, then at the next observation time, the number j in the system is given by

$$j = m + 1 - v, \qquad 0 \le v \le m + 1,$$

where v is the number of customers served between two successive arrivals.

Let $d_v(x)$ be the probability that v customers are served during time x. Since the number of services occur according to Poisson distribution with rate 1, then

$$d_v(x) = \frac{x^v e^{-x}}{v!}$$

The probability transition matrix $D(x)$ of the embedded Markov chain, $\{C_i, i \ge 0\}$, is determined by the probability that certain numbers of departures occurring during an interarrival time x,

$$D(x) = \begin{bmatrix} h_0(x) & d_0(x) & & & & \\ h_1(x) & d_1(x) & d_0(x) & & & \\ & & & \cdots & & \\ h_{N-2}(x) & d_{N-2}(x) & & & d_1(x) & d_0(x) \\ 0 & 0 & & & 0 & 1 \end{bmatrix}, \tag{9}$$

is an $N \times N$ matrix, where $h_i(x) = 1 - \sum_{j=0}^{i} d_j(x)$. The last row of $D(x)$ will never be used in computation. We have it in order to make a probability transition matrix.

Note that this embedded Markov chain is nonstationary. Equation (9) gives the transition probability matrix from time i to $i + 1$, which depends on the interarrival time x_i. Using this embedded Markov chain, we can derive the distribution functions of the customer delay times. We summarize the above discussion as the following proposition.

PROPOSITION 1: Given interarrival time $\bar{x}$, the state probability $\bar{p}_i$ and the individual expected delay time w_i can be calculated recursively as

$$\bar{p}_0 = u_k^T, \tag{10}$$

$$\bar{p}_i = \bar{p}_{i-1} D(x_i) \tag{11}$$

or

$$\bar{p}_i = \bar{p}_0 D(x_1) D(x_2) \ldots D(x_i), \tag{12}$$

and

$$w_i = \bar{p}_0 \left\{ \prod_{j=1}^{i} D(x_j) \right\} \beta, \qquad i = 1, 2, \ldots, n. \tag{13}$$

Equation (10) can be considered as the initial probability vector at time zero. Note that the kth element of $\bar{p}_0$ equals one representing k customers in the system initially. The expression for w_i only depends on x_j, for $j \leq i$. This is true because we have assumed that a first-come-first-served queueing discipline is employed. Therefore, only the first i interarrival times can affect the delay time of customer i. Equations (11) and (12) are similar to Chapman-Kolmogorov forward equations for a discrete-time Markov chain except that each probability transition matrix $D(x_i)$ varies on the basis of x_i in our case.

We now show that the expected customer delay time $w_i(\bar{x})$ is a decreasing convex function in $\bar{x}$. Direct algebraic manipulation is rather tedious. We show that the delay time $W_i(\bar{x})$ satisfies stochastic decreasing convexity in $\bar{x}$ instead. We need the following definition.

DEFINITION 2: The family of random variables $X(\theta)$ satisfies stochastic convexity (SCX) if for any θ, ρ there exist $Y(\theta) \stackrel{d}{=} X(\theta)$, $Y(\rho) \stackrel{d}{=} X(\rho)$, and $Y(\alpha\theta + (1 - \alpha)\rho) \stackrel{d}{=} X(\alpha\theta + (1 - \alpha)\rho)$ such that

$$Y(\alpha\theta + (1 - \alpha)\rho) \leq \alpha Y(\theta) + (1 - \alpha) Y(\rho)$$

almost surely for any $\alpha \in [1, 0]$. In addition, if for $\theta \leq \rho$, $Y(\theta) \geq Y(\rho)$ and $Y(\alpha\theta + (1 - \alpha)\rho) \geq Y(\rho)$ almost surely for any $\alpha \in [0, 1]$, then $X(\theta)$ satisfies stochastic decreasing convexity (SDCX).

Here $\stackrel{d}{=}$ means equal in distribution and θ is a parameter which is usually involved in the distribution function and can be either a scaler or a vector. Stochastic concavity and stochastic increasing convexity can be defined similarly.

From Shanthikumar and Yao [12], we collect some useful results as the following Lemma.

LEMMA 3: If $X(\theta) = \epsilon + f(\theta)$, where ϵ is a random component and f is an increasing/decreasing convex (deterministic) function in θ; then $X(\theta)$ satisfies stochastic increasing/decreasing convexity (SICX/SDCX). Furthermore, any operations of sum, maximum, or mixture on $X(\theta)$ result in SICX (SDCX).

It is not hard to obtain the following proposition by applying Lemma 3.

PROPOSITION 4: The customer delay time $W_i(\bar{x})$ satisfies stochastic decreasing convexity in $\bar{x}$.

PROOF: Denote S_i, $i = 1, 2, \ldots, N$, iid service time. Then

$$W_1 = \max\left\{\sum_{j=1}^{k} S_j - x_1, 0\right\} + S_{k+1}$$

and

$$W_{i+1} = \max\{W_i - x_{i+1}, 0\} + S_{k+i+1}.$$

Since $-x_i$ is a decreasing function and the recursive system only involves the operators of max and + and both are known to be SDCX [$-x_i$ can be considered as $+(-x_i)$], the result is immediate. ■

Since the delay time is decreasing convex, therefore, $J(\bar{x}, \bar{w})$ in (2) is convex in $\bar{x}$. The optimal $\bar{x}$ can be determined by differentiating $J(\bar{x}, \bar{w})$ by chain rule and solving n nonlinear resulting equations:

$$\left[\frac{dJ(\bar{x}, \bar{w})}{d\bar{x}}\right]^T = \left[\frac{\partial J(\bar{x}, \bar{w})}{\partial \bar{w}}\right]^T \frac{\partial \bar{w}^T}{\partial \bar{x}} + \left[\frac{\partial J(\bar{x}, \bar{w})}{\partial \bar{x}}\right]^T = \bar{0}, \tag{14}$$

where $\partial J(\bar{x}, \bar{w})/\partial \bar{w}$ and $\partial J(\bar{x}, \bar{w})/\partial \bar{x}$ are n-dimension vectors and

$$\frac{\partial \bar{w}^T}{\partial \bar{w}} = \left\{\frac{\partial w_i}{\partial x_j}\right\} \tag{15}$$

is an $n \times n$ lower triangular matrix whose elements can be derived as follows:

$$\text{for } i \geq j, \quad \frac{\partial w_i}{\partial x_j} = u_k^T \left\{\prod_{m=1}^{j-1} D(x_m)\right\} \frac{dD(x_j)}{dx_j} \left\{\prod_{m=j+1}^{i} D(x_m)\right\} \beta, \tag{16}$$

$$\text{for } i < j, \quad \frac{\partial w_i}{\partial x_j} = 0, \tag{17}$$

where the ill-defined product is evaluated to be the unit matrix I and

$$\frac{dD(x)}{dx} = \begin{bmatrix} d_0 & -d_0 & & & & \\ d_1 & d_0 - d_1 & -d_0 & & & \\ & & & \cdots & & \\ d_{N-1} & d_{N-2} - d_{N-1} & d_{N-3} - d_{N-2} & \cdots & d_0 - d_1 & -d_0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} \tag{18}$$

is directly computed from Eq. (9). The argument x is omitted for simplicity. Equation (14) can be solved by any of the nonlinear programming techniques [6] such as the gradient approach. After solving nonlinear equations (14) to get optimal values $\bar{x}^*$, we can compute the expected delay time w_i by the following procedure:

Step 1. $\bar{p} = u_k^T$, $w = 0$.
Step 2. For $i = 1, n$, do Step 3.
Step 3. Compute $D(x_i^*)$ by using Eq. (9), and

$$\bar{p} = \bar{p}D(x_i^*),$$

$$w_i = \bar{p}\beta,$$

and

$$w = w + w_i.$$

A Special Case: Constant Interarrival Times: If we restrict ourselves to constant interarrival times, i.e., $x_i = x$ for all i, $i = 1, 2, \ldots, n$, then the total delay time is simply a function of scalar x:

$$w(x) = u_k^T \sum_{i=1}^{n} D^i(x)\beta, \tag{19}$$

and the objective function becomes

$$J(x, \bar{w}(x)) = \alpha w(x) + (1 - \alpha)(nx + w_n(x)). \tag{20}$$

PROPOSITION 5: If x_n^* denotes the optimal interarrival time when n customers are scheduled and there is only one customer in the system at time zero ($k = 1$), we have the inequality

$$x_n^* \leq x_{n+1}^*, \tag{21}$$

and in addition,

$$x_1^* = -\ln(1 - \alpha) \tag{22}$$

$$\lim_{n\to\infty} x_n^* = 1 + \sqrt{\frac{\alpha}{2(1 - \alpha)}}. \tag{23}$$

PROOF: The proof of inequality (21) is straightforward by induction and is omitted. We only prove Eqs. (22) and (23).

For $n = 1$ the delay time $w_1(x) = e^{-x}$ and the objective function $J(x, w_1) = e^{-x} + (1 - \alpha)x$. An analytical value for x_1^* is easily obtained by solving $dJ/dx = 0$, which yields

$$x_1^* = -\ln(1 - \alpha).$$

When n approaches infinity, the problem becomes a $D/M/1$ queueing model and the objective function (20) is equivalent to

$$\text{minimize } J(x, w_q) = \alpha w_q + (1 - \alpha)x, \tag{24}$$

where w_q is the waiting time in a $D/M/1$-type queue. If the mean service rate is assumed to be one and the well-known upper bound of w_q in a $GI/G/1$ queue is used [13], then

$$w_q = \frac{1/x}{2(1 - 1/x)} = \frac{1}{2(x - 1)}. \tag{25}$$

Substituting (25) in (24), differentiating (24) with respect to x, and letting $dJ/dx = 0$ yields

$$\lim_{n\to\infty} x_n^* = 1 + \sqrt{\frac{\alpha}{2(1 - \alpha)}}.$$

■

In both cases, the bounds are functions of the weighting factor α only.

4. DYNAMIC SCHEDULING PROBLEMS

In static scheduling problems, we assumed that there is no scheduled customers to enter the system at time zero and we studied the n customer scheduling problem. In this section, we are concerned with a different scenario. Consider at any time, that there are k customers in the system already and there are n customers scheduled to come at time t_i, $i = 1, \ldots, n$, or $\bar{x}$. A new customer calls for an appointment. Then what is the optimal time for this new customer to come such that the objective function (2) still remains optimal?

We continue to use the notation employed in the previous section and need some new notation. We denote t for the arrival time of the new customer and $w(t)$ for the expected total delay time, $w(t) = \sum_{i=1}^{n+1} W_i(t)$. We assume that the old schedules cannot be changed but this new customer can be scheduled between any two customers. If $t_i < t < t_{i+1}$, we say the new customer is "inserted" between customer i and customer $i + 1$. If $t > t_n$, we say the new customer is scheduled after customer n. We will give the delay time distribution first, and then apply the optimization technique to obtain the optimal arrival time t.

Since the new customer can be inserted between any two customers, we partition the time interval $[0, \infty)$ into $n + 1$ subintervals by the n scheduled customer arrival times, i.e., $0 = t_0 \leq t_1 \leq t_2 \leq \cdots \leq t_n < \infty$. We derive the expressions for $w(t)$ in each subinterval $[t_{i-1}, t_i]$. For any arrival time t and schedule $\bar{x}$, we construct a new vector $\bar{z}(t)$ as follows:

$$i^* = \max_i \{i: t_i < t\},$$

$$z_i = x_i, \qquad \text{for } i \leq i^*,$$

$$z_{i^*+1} = t - t_{i^*},$$

$$z_{i^*+2} = t_{i^*+1} - t,$$

and

$$z_{i+1} = x_i, \qquad \text{for } i^* + 2 < i \leq n.$$

This new interarrival time vector $\bar{z}(t)$ is of order $(n + 1)$. It keeps the old schedule for the n customers but inserts the new customer between customer i^* and customer $i^* + 1$. Therefore, any pair $(\bar{x}, t)$ can be converted into $\bar{z}(t)$ and the problem can be expressed in terms of $\bar{z}(t)$ and called

(DYNAMIC)

$$\text{minimize } J[\bar{z}(t), \bar{w}(t)] = \alpha u^T \bar{w}(t) + (1 - \alpha)[u^T \bar{z}(t) + u_{n+1}^T w_{n+1}(t)],$$

$$\text{subject to } t \geq 0 \tag{26}$$

which can be evaluated the same way as we did before. The following proposition shows that the expected total delay is a continuous function of arrival time t.

PROPOSITION 6: The expected total delay time, $w(t)$, is a continuous function of the new customer's arrival time t.

PROOF: It is easy to verify that $w(t)$ is continuous for $t_{i-1} < t < t_i$, $1 \leq i \leq n$ and for $t > t_n$. So we only need to show that $w(t)$ is continuous at $t = t_i$, $1 \leq i \leq n$; i.e.,

$$\lim_{t \to t_i^-} w(t) = \lim_{t \to t_i^+} w(t).$$

From Eq. (9) we have

$$\lim_{x \to 0^+} D(x) = \begin{bmatrix} 0 & I \\ 0 & 1 \end{bmatrix} = G,$$

where I is a unit matrix of order $N - 1$. When $t \to t_i^-$, we have $z_i \to x_i$ and $z_{i+1} \to 0^+$,

$$\lim_{t \to t_i^-} w(t) = u_k^T \left\{ \sum_{j=1}^{n+1} \prod_{m=1}^{j} D(z_m) \right\} \beta$$

$$= u_k^T \left\{ \sum_{j=1}^{i} \prod_{m=1}^{j} D(z_m) + \prod_{j=1}^{i} D(z_j) G \sum_{j=i+2}^{n+1} \prod_{m=i+2}^{j} D(z_m) \right\} \beta.$$

Similarly, when $t \to t_i^+$, we have $z_{i+1} \to 0^+$ and $z_{i+2} \to x_{i+1}$,

$$\lim_{t \to t_i^+} w(t) = u_k^T \left(\sum_{j=1}^{i} \prod_{m=1}^{j} D(z_m) + \prod_{j=1}^{i} D(z_j) G \sum_{j=i+2}^{n+1} \prod_{m=i+2}^{j} D(z_m) \right) \beta.$$

Thus, $\lim_{t \to t_i^-} w(t) = \lim_{t \to t_i^+} w(t)$ and the proof is complete. ■

When t approaches t_i^-, this new customer is scheduled to come at the same time as customer i and is served before customer i. When t approaches t_i^+, this new customer is scheduled to come at the same time as customer i but is served after customer i. Since all of the customers are treated identically, interchanging service sequence will not affect the total delay time at all. The statement of above proposition is illustrated by Figure 1. The function $w(t)$ is a continuous function overall. Within each interval $[t_{i-1}, t_i]$, the function $w(t)$ is still a convex function, but overall, the function loses convexity, which complicates the optimization problem. We need to find a global minimum point. If denote the local

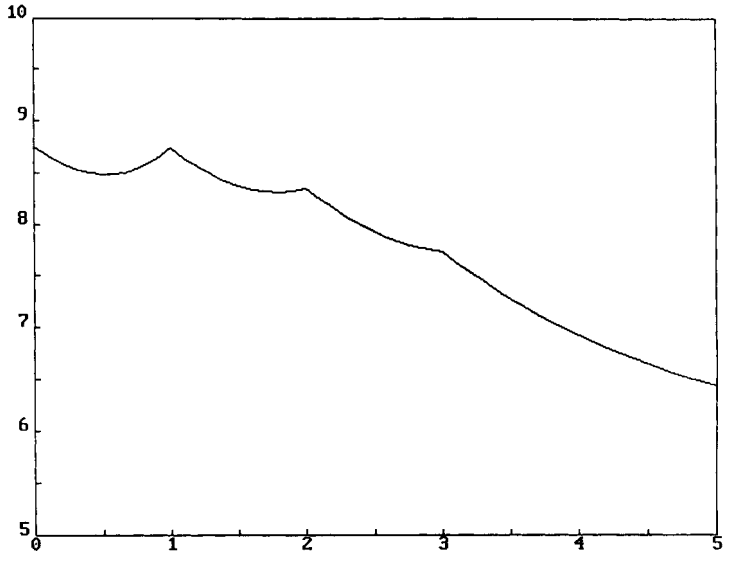

Figure 1.

minimum point $J_i^* = \min\{J[\bar{z}(t), \bar{w}(t)]; t_{i-1} \le t \le t_i\}$, then the global minimal J^* is just the smallest one among J_i^*, $i = 1, 2, \ldots, n+1$; i.e.,

$$J^* = \min_i\{J_i^*\}, \tag{27}$$

$$i^* = \arg\min_i\{J_i^*\}, \tag{28}$$

$$t^* = \arg\min\{J[\bar{z}(t), \bar{w}(t)]; \quad t_{i^*-1} \le t \le t_{i^*}\}. \tag{29}$$

Here, i^* is the index whose objective function has the smallest value and t^* is the optimal value which minimizes J_{i^*}.

5. NUMERICAL EXAMPLES

In the following examples, we minimize the combination of the expected total delay time and the completion time, which is

$$\text{minimize } \alpha u^T \bar{w} + (1 - \alpha)(u^T \bar{x} + u_n^T w_n),$$

$$\text{subject to } \bar{x} \ge 0.$$

In all of the examples below, we assume the service time has an exponential distribution with unit rate ($\mu = 1$) and initially there is one customer in the system ($k = 1$).

EXAMPLE 1 (static scheduling problems): For the simplest case $n = 1$, we have $x = -\ln(1 - \alpha)$. Analytic solutions for $\bar{x}$ do not exist for any $n > 1$. The simplest gradient method, that is, the fix-point algorithm, is employed in order to find the optimal interarrival times $\bar{x}$ as follows:

$$\bar{x}^{(m+1)} = \bar{x}^{(m)} - \gamma \frac{dJ[\bar{x}^{(m)}, \bar{w}(\bar{x}^{(m)})]}{d\bar{x}}, \tag{30}$$

where m is the number of iterations and γ is the step size. We set $\gamma = 1$ and start with $\bar{x}^{(0)} = u$.

Tables 1–4 show the optimal interarrival times for $n =$ two, three, four, and nine customers, respectively, with various values of the weighting factor α. Generally, the interarrival times increase when the weighting factor α increases. For a particular value of weighting factor α, the value of the interarrival time x_i increases when i is small and decreases when i is close to n. One may notice that the interarrival times are not equally spaced.

Table 1. Interarrival times for two customers.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_1	0.14	0.30	0.48	0.68	0.89	1.13	1.43	1.83	2.48
x_2	0.41	0.58	0.73	0.88	1.05	1.26	1.51	1.87	2.49

Table 2. Interarrival times for three customers.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_1	0.16	0.34	0.53	0.73	0.95	1.19	1.48	1.86	2.50
x_2	0.53	0.76	0.95	1.14	1.33	1.55	1.80	2.13	2.69
x_3	0.52	0.66	0.79	0.94	1.10	1.29	1.54	1.90	2.51

Table 3. Interarrival times for four customers.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_1	0.17	0.36	0.56	0.76	0.97	1.21	1.50	1.88	2.50
x_2	0.58	0.84	1.04	1.23	1.42	1.63	1.86	2.18	2.71
x_3	0.67	0.88	1.05	1.22	1.40	1.60	1.84	2.16	2.71
x_4	0.56	0.69	0.82	0.95	1.11	1.31	1.55	1.90	2.51

Table 4. Interarrival times for nine customers.

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_1	0.19	0.40	0.60	0.80	1.01	1.24	1.51	1.88	2.49
x_2	0.68	0.95	1.15	1.33	1.50	1.68	1.91	2.21	2.73
x_3	0.85	1.10	1.28	1.44	1.59	1.77	1.97	2.25	2.76
x_4	0.92	1.14	1.31	1.40	1.61	1.78	1.98	2.26	2.76
x_5	0.94	1.14	1.31	1.46	1.61	1.78	1.98	2.26	2.75
x_6	0.92	1.12	1.28	1.44	1.59	1.76	1.97	2.25	2.75
x_7	0.88	1.06	1.23	1.38	1.55	1.72	1.94	2.23	2.74
x_8	0.76	0.95	1.11	1.27	1.43	1.62	1.86	2.17	2.71
x_9	0.64	0.73	0.85	0.98	1.13	1.32	1.56	1.90	2.49

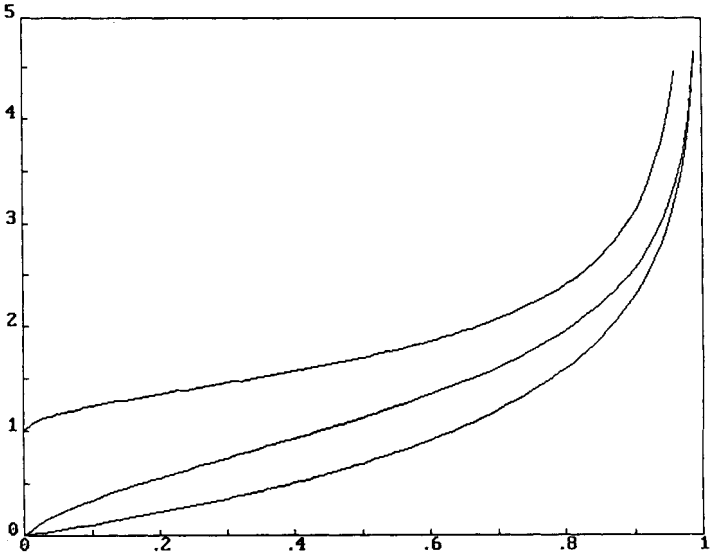

Figure 2.

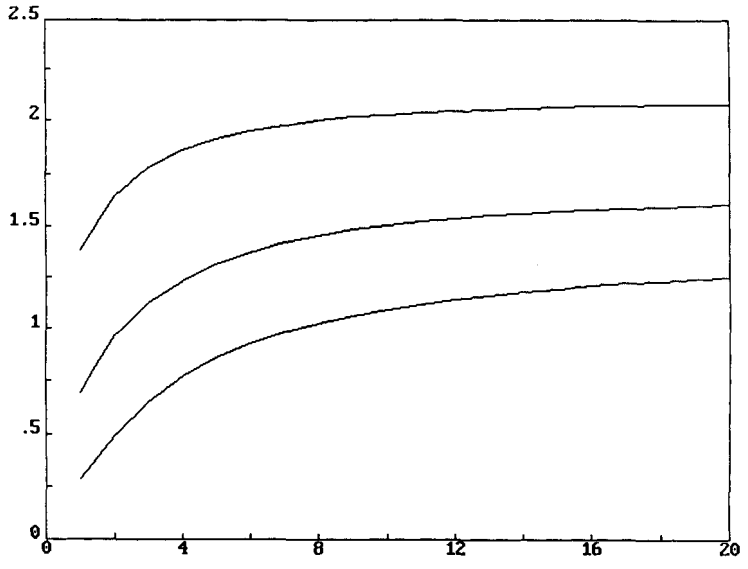

Figure 3.

EXAMPLE 2 [static scheduling problem (constant interarrival times)]: In this example, we consider constant interarrival times. Since the bounds are known by Proposition 5, the optimal interarrival times x_n^* can be determined by Fibonacci search. Figure 2 plots function x versus weighting factor α for $n = 4$. The top and bottom lines are the upper and lower bounds respectively. It shows that function x is a monotonic increasing function. It increases rapidly for smaller and larger α. In Figure 3, the optimal interarrival time x is plotted versus the

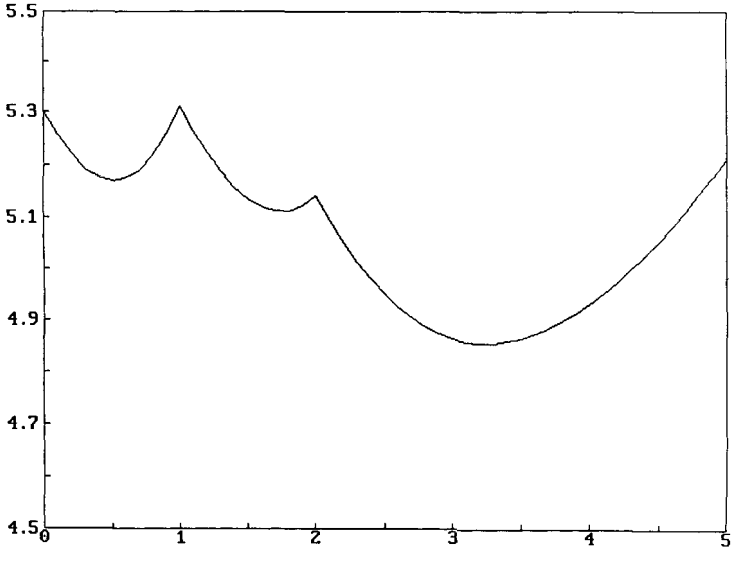

Figure 4.

number of customers to be scheduled, n, for various values of the weighting factor α. The top line represents the value of x_n^* for $\alpha = 0.75$, the middle line for $\alpha = 0.50$, and the bottom line for $\alpha = 0.25$. It shows that large values of α correspond to large values of interarrival time x_n^*.

EXAMPLE 3 (dynamic scheduling problem): In this example, we consider inserting a new customer when there are n scheduled customer in the system. Let $n = 2$, and $\bar{x} = [1, 1]$. In Figure 4, the objective function is plotted versus the arrival time of this new customer, t. The optimal arrival time occurs at $t = 3.2$, which is after customer 2.

6. SUMMARY AND EXTENSIONS

The optimal scheduling of a finite number of customer arrivals to a single-server system was studied in this article. The primary goal was to schedule the exact customer arrival times such that the weighted customer delay time and the system completion time is minimized. Two situations were analyzed. One is static scheduling: At time 0, there are k customers in the system and n customers are to be scheduled to come to the system. The other is dynamic scheduling: At time 0, there are k customers in the system and there are n customers scheduled to come to the system; only one customer is to be scheduled. In both situations, the customer delay time was shown to have a mixture of Erlang distributions which can be represented as a PH distribution. The expected delay time and its derivative are recursively computed. The optimal arrival times are then determined by solving a set of n nonlinear equations. Numerical results have shown that constant interarrival times cannot guarantee optimality.

The efficiency of the gradient method used here is under investigation [3], in which we have discovered that the convergency rate of $\bar{x}$ depends heavily on the weighting factor α. Smaller values of α correspond to faster convergency, and large values of α, close to one, correspond to slower convergency. Gradient methods other than the fixed point method are also studied.

The model developed here is valid only to a single-server system with exponential service times. It can be further investigated in different ways. First, since the phase-type distribution is matrix geometric in nature, the method can be modified to allow nonexponential service times. Second, the single queueing model can be extended to multiple servers, tandem queues, or, more generally, to queueing networks.

ACKNOWLEDGMENT

The author is grateful to Professor W. Gray, Mathematics Department of The University of Alabama, and two anonymous reviewers whose comments led to vast improvements in the quality of this manuscript.

REFERENCES

[1] Chaudhry, M., Harris, C., and Marchal, W., "Robustness of Rootfinding in Single-Server Queueing Models," *Journal of Computing*, **2**, 273–286 (1990).

[2] Grassmann, W., and Chaudhry, M., "A New Method to Solve Steady State Queueing Equations," *Naval Research Logistics Quarterly*, **29**(3), 461–473 (1982).

[3] Gray, W., Kakoli, B., and Wang, P.P., "On the Computation of Optimal Schedules in a Single-Server System," unpublished manuscript.
[4] Healy, K. J., "Scheduling Arrivals to a Stochastic Service Mechanism," unpublished manuscript.
[5] Jansson, B., "Choosing a Good Appointment System—A Study of Queues of the Type (D, M, 1)," *Operations Research*, 292–312 (1966).
[6] Luenberger, D.G., *Linear and Nonlinear Programming* (2nd ed.), Addison-Wesley, Reading, MA, 1984.
[7] Mercer, A., "A Queueing Problem in Which the Arrival Times of the Customers are Scheduled," *Journal of the Royal Statistical Society, Series B*, **22**, 108–116 (1960).
[8] Mercer, A., "Queues with Scheduled Arrivals. A Correction, Simplification, and Extension," *Journal of the Royal Statistical Society, Series B*, **35**, 104–116 (1973).
[9] Neuts, M.F., *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, Johns Hopkins University Press, Baltimore, MD, 1981.
[10] Rosenshine, M., "Scheduling Arrivals to Service Systems," presentation at ORSA/TIMS Joint National Metting, Las Vegas, 1990.
[11] Sabria, F., and Daganzo, C.F., "Approximate Expressions for Queueing Systems with Scheduled Arrivals and Established Service Order," *Transportation Science*, **23**(3), 159–165 (1989).
[12] Shanthikumar, J.G., and Yao, D.D., "Second-Order Stochastic Properties in Queueing Systems," *Proceedings of the IEEE, Special issue on Dynamics of Discrete Event Systems*, **77**, 162–170 (1989).
[13] Wolff, R.W., *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

Manuscript received January 15, 1992
Revised manuscript received July 17, 1992
Accepted September 21, 1992