Probability in the Engineering and Informational Sciences, 8, 1994, 491–509. Printed in the U.S.A.

OPTIMAL STOCHASTIC DYNAMIC SCHEDULING IN MULTI-CLASS QUEUES WITH TARDINESS AND/OR EARLINESS PENALTIES

DIMITRIOS G. PANDELIS AND DEMOSTHENIS TENEKETZIS
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, Michigan 48109-2122

Tasks belonging to N classes arrive for processing in a multi-server facility. Each arriving task has a due date (deterministic or random) associated with the completion of its service. If the service of a task is completed at a time other than the task's due date, an earliness or tardiness penalty is incurred. We determine properties of dynamic nonidling nonpreemptive and dynamic nonidling preemptive scheduling strategies that minimize an infinite horizon expected discounted cost due to the earliness and tardiness penalties. We provide examples that illustrate the properties of the optimal strategies.

1. INTRODUCTION – PROBLEM FORMULATION

We consider multi-server queueing systems where the tasks to be processed have constraints on their service completion times. Each arriving task belongs to one of N different classes and has a due date associated with the completion of its service; this due date either is random or becomes known at the arrival instant. Interarrival times, service times, and due dates form sequences of independent random variables that are also independent of each other. We assume that interarrival times have a general probability distribution.

This research was supported in part by the NSF under grant NCR-9204419.

Consider a task of type i, $i = 1,2,\ldots,N$, whose service completion time and due date are s and d, respectively. If $s < d$, an earliness penalty at a rate α_i per unit time is incurred from time s to time d. If $s \geq d$, a tardiness penalty at a rate β_i is incurred from time d to time s. We study problems with deterministic and stochastic due dates. We determine properties of dynamic nonidling nonpreemptive as well as nonidling preemptive scheduling strategies that minimize the infinite horizon expected γ-discounted cost due to the earliness and tardiness penalties.

Work on scheduling with tardiness and earliness penalties is motivated by production problems in manufacturing systems, where the emphasis is on Just-in-Time production (see Baker and Scudder [4]). In such a scheduling environment, both earliness and tardiness are penalized, because an early completion of a job may create the need for storage, thus incurring inventory costs, whereas a late completion may result in, for example, customer dissatisfaction or loss of sales. Applications involving tasks with priorities and due dates are also found in communication networks (see Bhattacharya and Ephremides [6]), where, for example, a single channel is used for the transmission of different types of messages (voice, video, data files, etc.) that have different priorities and constraints on the completion of their transmission, thus incurring tardiness costs when their due dates are not met.

Various versions of the deterministic scheduling problem with tardiness and/or earliness constraints are discussed in Baker and Scudder [4, and references therein], Szwarc [24], Du and Leung [8], Kubiak, Lou, and Sethi [15], Hall and Posner [11], Hall, Kubiak, and Sethi [10], and Potts and Van Wassenhove [22], where optimality and computational complexity issues are investigated and algorithms for the solution of these problems are proposed. Most of the research on stochastic scheduling with time constraints has concentrated on performance evaluation (e.g., Kaspi and Perry [13,14], Baccelli, Boyer, and Hebuterne [2], Baccelli and Trivedi [3], Perry and Levikson [20], and Bhattacharya and Ephremides [7]), rather than on optimization. For results on optimization see Panwar, Towsley, and Wolff [19], Bhattacharya and Ephremides [5], Frostig [9], and Pandelis and Teneketzis [18], where tasks have individual strict deadlines (i.e., the tasks incur zero penalty if their service begins or is completed before their due date, and they incur a fixed penalty depending on their priority otherwise), and Nain and Ross [16], Ross and Chen [23], and Altman and Shwartz [1], where there are constraints on the average delay of the tasks. For problems with tardiness penalties, Pinedo [21] and Huang and Weiss [12] study the optimality of the expected earliest due date service discipline when the due dates have a known probability distribution. Towsley and Baccelli [25] show the optimality of the earliest due date policy for a tandem network, where the due dates are known. Finally, Bhattacharya and Ephremides [6] consider a scheduling problem with tasks that belong to two different classes and show that the optimal policy is characterized by thresholds.

Our work deals with stochastic scheduling problems. It differs from Bhattacharya and Ephremides [6] and Pinedo [21] in that it deals with many servers and from Huang and Weiss [12] and Towsley and Baccelli [25] in that it deals with many classes of tasks; furthermore, our work incorporates earliness penalties.

The contributions of this work are the following: we characterize qualitative properties of optimal dynamic nonidling nonpreemptive as well as of optimal dynamic nonidling preemptive scheduling strategies for the single- and multiserver problem with tardiness and/or earliness penalties. We provide examples that illustrate the properties of the optimal strategies.

The remainder of the paper is organized as follows: In Section 2 we consider nonpreemptive policies. In Section 3 we consider preemptive policies, where preemptions are allowed at service completions and arrival times.

2. NONPREEMPTIVE POLICIES

2.1. Deterministic Due Dates

We consider a queueing system with M identical parallel servers $S_1, S_2, \ldots, S_M$, $M \geq 1$. In the case of nonpreemptive policies decisions have to be made at service completion times.

Let $t_0, t_0 \geq 0$, be such a decision point and $M^i(t_0)$, $i = 1,2,\ldots,N$, be the set of tasks of type i that are present in the system at time t_0. From the decision maker's point of view, a task with due date less than t_0 is equivalent to a task with due date equal to t_0, because for any scheduling strategy and any realization of the arrival, service, and due dates processes the two tasks incur the same cost after time t_0. Adopting the convention that tasks with due dates less than t_0 have due dates equal to t_0, we define for $M^i(t_0) \neq \emptyset$ $D^i(t_0) = \{d_1^i, \ldots, d_{n_i}^i\}$, $n_i \geq 1$, to be the set of due dates of tasks of type i present in the system at time t_0 arranged in increasing order. From now on d_j^i will denote both the time instant d_j^i and the task that has due date d_j^i.

We assume that service times for tasks in class i, $i = 1,2,\ldots,N$, have a general probability distribution denoted by F_i. For a decision instant t_0, if M', $M' \leq M$, is the number of empty servers, the problem is to determine the M' tasks to be processed. The following theorem describes the characteristics of the optimal policy.

THEOREM 1: *Let t_0 be a decision instant for which servers $S_{n_1}, S_{n_2}, \ldots S_{n_{M'}}$, $M' \leq M$, are idle. Consider server S_{n_ℓ}, $1 \leq \ell \leq M'$. Then:*

(i) *Within a class of tasks, it is optimal to assign to S_{n_ℓ} the one with the earliest due date.*

(ii) *For each pair of classes i,j ($i,j = 1,2,\ldots,N, i \neq j$) with $M^i(t_0) \neq \emptyset$, $M^j(t_0) \neq \emptyset$, there exists a time instant (threshold) t_{ij} such that it is* *optimal to assign to S_{n_ℓ} task d_1^i instead of d_1^j if $d_1^j \geq t_{ij}$, and vice versa otherwise.*

PROOF:

(i) Consider a class i of tasks with at least two tasks waiting to be processed at time t_0. For every policy π that assigns to S_{n_ℓ} task $d_k^i, k \neq 1$, we construct policy $\tilde{\pi}$ as follows: a time t_0 $\tilde{\pi}$ assigns to S_{n_ℓ} task d_1^i and is identical to π when π processes tasks other than d_1^i. When π processes (if ever) task d_1^i at some server, $\tilde{\pi}$ processes task d_k^i at the same server. Let V^π and $V^{\tilde{\pi}}$ be the costs incurred under policies π and $\tilde{\pi}$, respectively. For any sample path all tasks except d_1^i, d_k^i incur the same cost along π and $\tilde{\pi}$. Let σ' be the time π starts processing d_1^i ($\sigma' = \infty$ if π never processes d_1^i) and σ and τ be the processing times of tasks d_1^i and d_k^i (respectively, d_k^i and d_1^i) under π (respectively, $\tilde{\pi}$). Then,

$$E(V^\pi - V^{\tilde{\pi}} \,|\, \sigma' = \ell) = E[C^i(d_k^i, t_0 + \tau) + C^i(d_1^i, \ell + \sigma) - C^i(d_1^i, t_0 + \tau) - C^i(d_k^i, \ell + \sigma)], \qquad (1)$$

where $C^i(d,s)$ is the cost incurred by a task of class i with due date d and service completion time s. We have

$$C^i(d,s) = \alpha_i e^{-\gamma s} \int_0^{(d-s)^+} e^{-\gamma t}\, dt + \beta_i e^{-\gamma d} \int_0^{(s-d)^+} e^{-\gamma t}\, dt = \frac{1}{\gamma}\,[\alpha_i e^{-\gamma s} + \beta_i e^{-\gamma d} - (\alpha_i + \beta_i) e^{-\gamma \max\{d,s\}}]. \qquad (2)$$

From Eqs. (1) and (2), we get

$$E(V^\pi - V^{\tilde{\pi}} \,|\, \sigma' = \ell) = \frac{\alpha_i + \beta_i}{\gamma}\, E[e^{-\gamma \max\{d_1^i, t_0+\tau\}} + e^{-\gamma \max\{d_k^i, \ell+\sigma\}} - e^{-\gamma \max\{d_k^i, t_0+\tau\}} - e^{-\gamma \max\{d_1^i, \ell+\sigma\}}].$$

Since σ and τ have the same probability distribution F_i, the preceding expectation will not change if we replace them with a random variable f_i that has distribution F_i. Therefore,

$$E(V^\pi - V^{\tilde{\pi}} \,|\, \sigma' = \ell) = \frac{\alpha_i + \beta_i}{\gamma}\, E[e^{-\gamma \max\{d_1^i, t_0+f_i\}} + e^{-\gamma \max\{d_k^i, \ell+f_i\}} - e^{-\gamma \max\{d_k^i, t_0+f_i\}} - e^{-\gamma \max\{d_1^i, \ell+f_i\}}].$$

The expression over which the expectation is taken is nonnegative for $d_1^i < d_k^i$ and $t_0 \leq \ell$. Therefore among all tasks of type i it is optimal to process the one with the earliest due date.

(ii) Let d^i and d^j be tasks of type i and j that are present in the system at time t_0. Let the due dates of all tasks except d^j be fixed. It suffices to show that if for $d^j = k$ it is optimal to process d^i instead of d^j, then for $d^j = \ell$, where $\ell > k$, it is still optimal to process d^i.

To avoid confusion we will attach to each policy a subscript denoting the value of d^j. For example, π_k would be a policy applied to the set of tasks with $d^j = k$.

Let π_ℓ be a policy that processes task d^j at time t_0 and π_k a policy identical to π_ℓ. Let also σ be the service time of task d^j. Then we have from Eq. (2)

$$\begin{aligned} E(V^{\pi_k} - V^{\pi_\ell}) &= E[C^j(k, t_0 + \sigma) - C^j(\ell, t_0 + \sigma)] \\ &= \frac{1}{\gamma} E[\beta_j(e^{-\gamma k} - e^{-\gamma \ell}) \\ &\quad - (\alpha_j + \beta_j)(e^{-\gamma \max\{k, t_0+\sigma\}} - e^{-\gamma \max\{\ell, t_0+\sigma\}})]. \qquad (3) \end{aligned}$$

By assumption there exists a policy $\tilde{\pi}_k$ that processes task d^i at t_0 and does better than π_k; that is,

$$E[V^{\pi_k} - V^{\tilde{\pi}_k}) \geq 0. \qquad (4)$$

Let $\tilde{\pi}_\ell$ be a policy identical to $\tilde{\pi}_k$ and τ be the time the service of task d^j starts under $\tilde{\pi}_k, \tilde{\pi}_\ell$. Then we have from Eq. (2)

$$\begin{aligned} E(V^{\tilde{\pi}_k} - V^{\tilde{\pi}_\ell} \mid \tau = \tau') &= E[C^j(k, \tau' + \sigma) - C^j(\ell, \tau' + \sigma)] \\ &= \frac{1}{\gamma} E[\beta_j(e^{-\gamma k} - e^{-\gamma \ell}) - (\alpha_j + \beta_j) \\ &\quad \times (e^{-\gamma \max\{k, \tau'+\sigma\}} - e^{-\gamma \max\{\ell, \tau'+\sigma\}})]. \qquad (5) \end{aligned}$$

It is straightforward to show that for $t_0 \leq \tau'$ the right-hand side (RHS) of Eq. (3) is less than or equal to the RHS of Eq. (5). Therefore,

$$\begin{aligned} &E(V^{\pi_k} - V^{\pi_\ell}) \leq E(V^{\tilde{\pi}_k} - V^{\tilde{\pi}_\ell}) \\ &\Rightarrow E(V^{\pi_\ell} - V^{\tilde{\pi}_\ell}) \geq E(V^{\pi_k} - V^{\tilde{\pi}_k}) \geq 0, \end{aligned}$$

where the last inequality follows from Eq. (4). Therefore, it is optimal to process d^i when $d^j = \ell$. ■

Remark 1: If the service times for classes i and j are identically distributed, and $\alpha_i \leq \alpha_j$, $\beta_i \geq \beta_j$, then the threshold t_{ij} in part (ii) of Theorem 1 is less than d_1^i.

According to Theorem 1 and the preceding remark, to determine which task is optimal to process we need to make at most $N-1$ pairwise comparisons—in other words, compute at most $N-1$ thresholds. The selection of the M' tasks to be processed is done sequentially. First, we determine the "best" task by computing the appropriate thresholds; then we determine the "best" among the rest of the tasks, and so on.

Remark 2: Note that the threshold t_{ij} (in part (ii) of Theorem 1) may not only depend on d_1^i, d_1^j but on the whole set of tasks present in the system as well as on the number of servers. These points are illustrated by examples in Pandelis [17].

2.2. Stochastic Due Dates

Contrary to the problem examined in Section 2.1, the tasks' due dates do not become known at the arrival instants. Consider a task of type $i, i = 1,2,\ldots,N$, that arrives at time ρ. Its due date is equal to $\rho + X_i$, where X_i is a random variable with probability distribution function denoted by G_i. We assume that a task's due date becomes known at the time of its service completion.

For a decision instant t_0, we denote by $M^i(t_0), i = 1,2,\ldots,N$, the set of tasks of type i that are present in the system at time t_0. For $M^i(t_0) \neq \emptyset$, we define $A^i(t_0) = \{a_1^i,\ldots,a_{n_i}^i\}$, $n_i \geq 1$, to be the set of arrival times of tasks of type i present in the system at time t_0 arranged in increasing order. From now on a_j^i will denote both the time instant a_j^i and the task with arrival time a_j^i.

We assume that service times for tasks of type i, $i = 1,2,\ldots,N$, have a general probability distribution denoted by F_i. The properties of an optimal policy are given in the following theorem.

Theorem 2: *Let t_0 be a decision instant for which servers $S_{n_1}, S_{n_2}, \ldots, S_{n_{M'}}$, $M' \leq M$, are idle. Consider server S_{n_ℓ}, $1 \leq \ell \leq M'$. Then:*

(i) *Within a class of tasks it is optimal to assign to S_{n_ℓ} the one with the earliest arrival time.*

(ii) *For each pair of classes i,j $(i,j = 1,2,\ldots,N,\ i \neq j)$ with $M^i(t_0) \neq \emptyset$, $M^j(t_0) \neq \emptyset$, there exists a time instant (threshold) t_{ij} such that it is optimal to assign to S_{n_ℓ} task a_1^i instead of a_1^j if $a_1^j \geq t_{ij}$, and vice versa otherwise.*

Proof:

(i) Consider a class i of tasks with at least two tasks waiting to be processed at time t_0. For every policy π that assigns to S_{n_ℓ} task $a_k^i, k \neq 1$, we construct policy $\tilde{\pi}$ as follows: at time t_0 $\tilde{\pi}$ assigns to S_{n_ℓ} task a_1^i and is identical to π when π processes tasks other than a_1^i. When π processes (if ever) task a_1^i at some server, $\tilde{\pi}$ processes task a_k^i at the same server. The due dates of a_1^i and a_k^i are $a_1^i + x_1^i$ and $a_k^i + x_k^i$, respectively, where x_1^i and x_k^i are random variables with distribution G_i. Let σ' be the time π starts processing a_1^i ($\sigma' = \infty$ if π never processes a_1^i) and σ, τ be the processing times of tasks a_1^i, a_k^i under π (respectively, $\tilde{\pi}$). Then,

$$\begin{aligned} E(V^{\pi} - V^{\tilde{\pi}} \,|\, \sigma' = \ell) &= E[C^i(a_k^i + x_k^i, t_0 + \tau, t_0) + C^i(a_1^i + x_1^i, \ell + \sigma, t_0) \\ &\qquad - C^i(a_1^i + x_1^i, t_0 + \tau, t_0) - C^i(a_k^i + x_k^i, \ell + \sigma, t_0)], \end{aligned} \tag{6}$$

where $C^i(d, s, t_0)$ is the cost incurred after time t_0 by a task of class i with due date d and service completion time s. We have

$$\begin{aligned} C^i(d, s, t_0) &= \alpha_i e^{-\gamma s} \int_0^{(d-s)^+} e^{-\gamma t}\, dt \\ &\quad + \beta_i e^{-\gamma \max\{d, t_0\}} \int_0^{(s - \max\{d, t_0\})^+} e^{-\gamma t}\, dt \\ &= \frac{1}{\gamma} [\alpha_i e^{-\gamma s} + \beta_i e^{-\gamma \max\{d, t_0\}} - (\alpha_i + \beta_i) e^{-\gamma \max\{d, s\}}]. \end{aligned} \tag{7}$$

From Eqs. (6) and (7), we get

$$\begin{aligned} E(V^{\pi} - V^{\tilde{\pi}} \,|\, \sigma' = \ell) &= \frac{\alpha_i + \beta_i}{\gamma} E[e^{-\gamma \max\{a_1^i + x_1^i, t_0 + \tau\}} + e^{-\gamma \max\{a_k^i + x_k^i, \ell + \sigma\}} \\ &\qquad - e^{-\gamma \max\{a_k^i + x_k^i, t_0 + \tau\}} - e^{-\gamma \max\{a_1^i + x_1^i, \ell + \sigma\}}]. \end{aligned}$$

The preceding expectation will not change if we replace σ, τ with a random variable f_i having distribution F_i, and x_1^i, x_k^i with a random variable x_i having distribution G_i. Therefore,

$$\begin{aligned} E(V^{\pi} - V^{\tilde{\pi}} \,|\, \sigma' = \ell) &= \frac{\alpha_i + \beta_i}{\gamma} E[e^{-\gamma \max\{a_1^i + x_i, t_0 + f_i\}} + e^{-\gamma \max\{a_k^i + x_i, \ell + f_i\}} \\ &\qquad - e^{-\gamma \max\{a_k^i + x_i, t_0 + f_i\}} - e^{-\gamma \max\{a_1^i + x_i, \ell + f_i\}}]. \end{aligned}$$

The expression over which the expectation is taken is nonnegative for $a_1^i < a_k^i$ and $t_0 \leq \ell$.

Therefore among all tasks of type i it is optimal to process the one with the earliest arrival time.

(ii) Let a^i and a^j be tasks of type i and j that are present in the system at time t_0. Let the arrival times of all tasks except a^j be fixed. It suffices to show that if for $a^j = k$ it is optimal to process a^i instead of a^j, then for $a^j = \ell$, where $\ell > k$, it is still optimal to process a^i.

For a specific sample path, the due date of a^j when $a^j = k$ and $a^j = \ell$ is $k + x_j$ and $\ell + x_j$, respectively, where x_j is the realization of the random variable X_j for this sample path. Since $k + x_j < \ell + x_j$, we can use the arguments in the proof of part (ii) of Theorem 1 to prove the result. ■

Remark 3: If the service times and the due dates for classes i and j are identically distributed, and $\alpha_i \le \alpha_j, \beta_i \ge \beta_j$, then the threshold t_{ij} in part (ii) of Theorem 2 is less than a_1^i.

It is interesting to note the similarity between Theorems 1 and 2. By replacing due dates with arrival times, we get the statement of Theorem 2 from that of Theorem 1. Both theorems combined with Remarks 1 and 3 assert that the tasks that are optimal to process can be sequentially determined, each one by at most $N - 1$ pairwise comparisons.

3. PREEMPTIVE POLICIES

3.1. Deterministic Due Dates

We consider a queueing system with M parallel exponential servers $S_1, S_2, \ldots, S_M$, with rates $\mu_1, \mu_2, \ldots, \mu_M$, respectively. The assumption of exponential service times is crucial when we consider preemptive policies. Because we consider preemptive policies, at any event, arrival, or service completion, a decision has to be made on how to assign the tasks present in the system to the M servers. We investigate the multi-server allocation problem by considering two cases: (i) when the number of tasks is greater than or equal to the number of servers, and (ii) when the number of tasks is smaller than the number of servers.

Consider first a decision instant for which the number of tasks is greater than or equal to the number of servers. Properties of the optimal policy are described in the following theorem.

THEOREM 3:

(i) *Let t_0 be a decision instant for which the number of tasks is greater than or equal to the number of servers and S_j be the fastest server. Then, within a class of tasks it is optimal to assign to S_j the one with the earliest due date.*

(ii) *In the single-server problem ($M = 1$), for each pair of classes i, j ($i, j = 1, 2, \ldots, N$, $i \neq j$) with $M^i(t_0) \neq \emptyset$, $M^j(t_0) \neq \emptyset$, there exists a time instant (threshold) t_{ij} such that it is optimal to process task d_1^i instead of d_1^j if $d_1^j \ge t_{ij}$, and vice versa otherwise.*

PROOF:

(i) Consider a class i of tasks with at least two tasks present in the system at time t_0. Let π be a policy that at time t_0 assigns to S_j task $d_k^i, k \neq 1$, and proceeds optimally after t_0. We construct policy $\tilde{\pi}$ as follows: at time t_0 $\tilde{\pi}$ assigns tasks in the same way as π, except that it assigns task d_l^i to S_j and, if π has assigned d_l^i to some other server, it assigns task d_k^i to the same server. After time t_0 $\tilde{\pi}$ proceeds optimally. We also construct policy π' as follows: at time t_0 π' assigns task d_l^i to S_j and is identical to π when π processes tasks other than d_l^i. When π processes (if ever) task d_l^i at some server, π' processes task d_k^i at the same server.

We consider first the case when π assigns d_l^i to some server S_ℓ at t_0. When the first event after time t_0 is an arrival or a service completion at some server other than S_j or S_ℓ, then, because of the exponentiality of the service times, policies π and $\tilde{\pi}$ are coupled, thus incurring the same costs. With Ω being the basic sample space, we define events A^σ, Σ_j^σ, and Σ_ℓ^σ as follows:

$$A^\sigma = \{\omega \in \Omega \mid \text{first arrival after } t_0 \text{ occurs after time } \sigma\},$$
$$\Sigma_j^\sigma = \{\omega \in \Omega \mid \text{first event after } t_0 \text{ is service completion at server } S_j \text{ at time } \sigma\},$$
$$\Sigma_\ell^\sigma = \{\omega \in \Omega \mid \text{first event after } t_0 \text{ is service completion at server } S_\ell \text{ at time } \sigma\}.$$

Then, we have

$$\begin{aligned} E[V^\pi - V^{\tilde{\pi}}) = {} & \int_{t_0}^{\infty} E(V^\pi - V^{\tilde{\pi}} \mid \Sigma_j^\sigma) P(A^\sigma) \Bigl[\prod_{m \neq j,\ell} e^{-\mu_m(\sigma - t_0)} \Bigr] \\ & \times e^{-\mu_\ell(\sigma - t_0)} \mu_j e^{-\mu_j(\sigma - t_0)} \, d\sigma \\ & + \int_{t_0}^{\infty} E(V^\pi - V^{\tilde{\pi}} \mid \Sigma_\ell^\sigma) P(A^\sigma) \Bigl[\prod_{m \neq j,\ell} e^{-\mu_m(\sigma - t_0)} \Bigr] \\ & \times e^{-\mu_j(\sigma - t_0)} \mu_\ell e^{-\mu_\ell(\sigma - t_0)} \, d\sigma \\ = {} & \int_{t_0}^{\infty} E(V_k^\sigma - V_l^\sigma) P(A^\sigma) \prod_m e^{-\mu_m(\sigma - t_0)} (\mu_j - \mu_\ell) \, d\sigma, \quad \textbf{(8)} \end{aligned}$$

where V_k^σ (respectively, V_l^σ) is the cost incurred by the optimal policy, denoted by π_k^* (respectively, π_l^*), after time σ given that task d_k^i (respectively, d_l^i) completes its service at time σ and no other event has occurred until time σ. We construct policy $\tilde{\pi}_l^*$ as follows: $\tilde{\pi}_l^*$ is identical to π_k^* when π_k^* processes tasks other than d_l^i. When π_k^* processes (if ever) task d_l^i, $\tilde{\pi}_l^*$ processes task d_k^i. Let τ be the service completion time of d_l^i and d_k^i under π_k^* and $\tilde{\pi}_l^*$, respectively. We take $\tau = \infty$ if π_k^* never processes d_l^i. Let $V^{\tilde{\pi}_l^*}$ be the cost incurred under policy $\tilde{\pi}_l^*$ after time σ. Then,

$$\begin{aligned} V_k^\sigma - V^{\tilde{\pi}_l^*} = {} & C^i(d_k^i, \sigma, \sigma) + C^i(d_l^i, \tau, \sigma) \\ & - C^i(d_l^i, \sigma, \sigma) - C^i(d_k^i, \tau, \sigma). \qquad \textbf{(9)} \end{aligned}$$

From Eqs. (9) and (7), we get

$$V_k^\sigma - V^{\tilde{\pi}_1^*} = \frac{\alpha_i + \beta_i}{\gamma} [e^{-\gamma \max\{d_1^i, \sigma\}} + e^{-\gamma \max\{d_k^i, \tau\}} - e^{-\gamma \max\{d_k^i, \sigma\}} - e^{-\gamma \max\{d_1^i, \tau\}}],$$

which is a nonnegative expression for $d_1^i < d_k^i$ and $\sigma \leq \tau$. Therefore,

$$E(V_k^\sigma) \geq E(V^{\tilde{\pi}_1^*}) \geq E(V_1^\sigma), \tag{10}$$

where the second inequality follows from the fact that policy π_1^* is optimal, while $\tilde{\pi}_1^*$ is not necessarily optimal. From Eqs. (8) and (10) and $\mu_j \geq \mu_\ell$, we conclude that $E(V^\pi) \geq E(V^{\tilde{\pi}})$.

We now consider the case when π does not process d_1^i at t_0. We define events A_j and Σ_j as follows:

$A_j = \{\omega \in \Omega \,|\,$ first event after t_0 is an arrival or a service completion at a server other than $S_j\}$,

$\Sigma_j = \{\omega \in \Omega \,|\,$ first event after t_0 is service completion at server $S_j\}$.

Using the same interchange argument as in the proof of part (i) of Theorem 1, we get

$$E(V^{\pi'} | \Sigma_j) \leq E(V^\pi | \Sigma_j). \tag{11}$$

Moreover, because $\tilde{\pi}$ acts optimally after t_0, while π' does not necessarily do so, we have

$$E(V^{\tilde{\pi}} | \Sigma_j) \leq E(V^{\pi'} | \Sigma_j). \tag{12}$$

Finally, because of the exponentiality of the service times, we have

$$E(V^\pi | A_j) = E(V^{\tilde{\pi}} | A_j). \tag{13}$$

From Eqs. (11)-(13), it follows that $E(V^\pi) \geq E(V^{\tilde{\pi}})$.

Therefore, among all tasks of type i it is optimal to assign to S_j the one with the earliest due date.

(ii) Let d^i and d^j be tasks of type i and j that are present in the system at time t_0. Let the due dates of all tasks except d^j be fixed. It suffices to show that if for $d^j = k$ it is optimal to process d^i instead of d^j, then for $d^j = \ell$, where $\ell > k$, it is still optimal to process d^i.

As in the proof of part (ii) of Theorem 1, we attach to each policy a subscript denoting the value of d^j. Let π_k, π_ℓ be the policies that process task d^j at time t_0 and proceed optimally afterward, $\tilde{\pi}_k, \tilde{\pi}_\ell$ the policies that process task d^i at time t_0 and proceed optimally afterward, and π'_ℓ a policy that is identical to $\tilde{\pi}_k$. For sample paths for which the first event after t_0 is an arrival, policies $\pi_k, \tilde{\pi}_k$ and $\pi_\ell, \tilde{\pi}_\ell$ are coupled at the time of the arrival because of the exponentiality of the service times. Therefore,

$$E(V^{\pi_k} - V^{\tilde{\pi}_k} | A) = E(V^{\pi_\ell} - V^{\tilde{\pi}_\ell} | A) = 0. \qquad \textbf{(14)}$$

We consider now sample paths for which the first event after t_0 is a service completion. Let σ be the time of the first service completion and σ' the time the service of d^j is completed under policies $\tilde{\pi}_k, \pi'_\ell$. We have from Eq. (2)

$$\begin{aligned} V^{\pi_k} - V^{\pi_\ell} &= C^j(k,\sigma) - C^j(\ell,\sigma) \\ &= \frac{1}{\gamma}\,[\beta_j(e^{-\gamma k} - e^{-\gamma \ell}) - (\alpha_j + \beta_j) \\ &\qquad \times (e^{-\gamma \max\{k,\sigma\}} - e^{-\gamma \max\{\ell,\sigma\}})] \end{aligned} \qquad \textbf{(15)}$$

and

$$\begin{aligned} V^{\tilde{\pi}_k} - V^{\pi'_\ell} &= C^j(k,\sigma') - C^j(\ell,\sigma') \\ &= \frac{1}{\gamma}\,[\beta_j(e^{-\gamma k} - e^{-\gamma \ell}) - (\alpha_j + \beta_j) \\ &\qquad \times (e^{-\gamma \max\{\mathrm{k},\sigma'\}} - e^{-\gamma \max\{\ell,\sigma'\}})]. \end{aligned} \qquad \textbf{(16)}$$

From Eqs. (15) and (16), we get for $k < \ell$ and $\sigma \le \sigma'$

$$E(V^{\pi_k} - V^{\pi_\ell} | \Sigma) \le E(V^{\tilde{\pi}_k} - V^{\pi'_\ell} | \Sigma). \qquad \textbf{(17)}$$

Moreover, because $\tilde{\pi}_\ell$ acts optimally after the first service completion, while π'_ℓ does not necessarily do so, we have

$$E(V^{\tilde{\pi}_k} - V^{\pi'_\ell} | \Sigma) \le E(V^{\tilde{\pi}_k} - V^{\tilde{\pi}_\ell} | \Sigma). \qquad \textbf{(18)}$$

From Eqs. (14), (17), and (18), we get

$$\begin{aligned} E(V^{\pi_\ell} - V^{\tilde{\pi}_\ell}) &= E(V^{\pi_\ell} - V^{\tilde{\pi}_\ell} | \Sigma)P(\Sigma) \\ &\ge E(V^{\pi_k} - V^{\tilde{\pi}_k} | \Sigma)P(\Sigma) = E(V^{\pi_k} - V^{\tilde{\pi}_k}) \\ &\ge 0, \end{aligned}$$

where the last inequality follows from the assumption that d^i is optimal when $d^j = k$. Therefore, d^i is optimal when $d^j = \ell$. ■

Next, we consider decision instants for which the number of tasks is smaller than the number of servers. In such a situation the problem is to determine which servers will process the tasks and how to allocate the tasks to these servers. We restrict attention to list scheduling strategies, that is, strategies that select the servers according to a prespecified priority rule $(S_{\ell_1}, S_{\ell_2}, \ldots, S_{\ell_M})$, where

$(\ell_1, \ell_2, \ldots, \ell_M)$ is a permutation of $(1,2,\ldots,M)$. The following theorem describes properties of the optimal policy.

THEOREM 4: *Assume that we have only one class of tasks and only tardiness penalties are incurred. Let t_0 be a decision instant for which the number of tasks is smaller than the number of servers. Then:*

(i) *Once the servers have been selected, it is optimal to assign the task with the earliest due date to the fastest server.*

(ii) *The optimal list scheduling policy ranks the servers in decreasing order of their service rates, that is, uses the fastest servers.*

PROOF:

(i) The proof uses the same arguments as the proof of part (i) of Theorem 3.

(ii) Let π be a policy that follows the priority rule $(S_{\ell_1}, S_{\ell_2}, \ldots, S_{\ell_M})$ and $i = \min\{j : \mu_{\ell_j} < \mu_{\ell_{j+1}}\}$. Let τ_0 be the first time π selects S_{ℓ_i} instead of $S_{\ell_{i+1}}$. We construct policy π^1 as follows: π^1 is identical to π until τ_0, selects $S_{\ell_{i+1}}$ instead of S_{ℓ_i} at τ_0, and follows the same priority rule as π afterward.

Let $\sigma_j, j = 1,2,\ldots,i+1$, be the service completion times at servers $S_{\ell_1}, S_{\ell_2}, \ldots, S_{\ell_{i+1}}$, and ρ be the time of the first arrival after τ_0. Let $\tau = \min\{\rho, \sigma_1, \sigma_2, \ldots, \sigma_{i-1}\}$ and F its probability distribution. Then, because of the exponentiality of the service times, we have

$$E(V^\pi - V^{\pi^1} \,|\, \tau < \sigma_i, \sigma_{i+1}) = 0.$$

Therefore,

$$\begin{aligned}
&E(V^\pi - V^{\pi^1}) \\
&\quad = \int_{\tau_0}^{\infty} \int_{m}^{\infty} E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \sigma_i = n, \tau \geq n) \\
&\qquad \times \mu_{\ell_{i+1}} e^{-\mu_{\ell_{i+1}}(m-\tau_0)} \mu_{\ell_i} e^{-\mu_{\ell_i}(n-\tau_0)} [1 - F(n)] \, dn \, dm \\
&\qquad + \int_{\tau_0}^{\infty} \int_{m}^{\infty} E(V^\pi - V^{\pi^1} \,|\, \sigma_i = m, \sigma_{i+1} = n, \tau \geq n) \\
&\qquad \times \mu_{\ell_i} e^{-\mu_{\ell_i}(m-\tau_0)} \mu_{\ell_{i+1}} e^{-\mu_{\ell_{i+1}}(n-\tau_0)} [1 - F(n)] \, dn \, dm \\
&\qquad + \int_{\tau_0}^{\infty} \int_{m}^{\infty} E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \tau = n, \sigma_i \geq n) \\
&\qquad \times \mu_{\ell_{i+1}} e^{-\mu_{\ell_{i+1}}(m-\tau_0)} e^{-\mu_{\ell_i}(n-t_0)} \, dF(n) \, dm \\
&\qquad + \int_{\tau_0}^{\infty} \int_{m}^{\infty} E(V^\pi - V^{\pi^1} \,|\, \sigma_i = m, \tau = n, \sigma_{i+1} \geq n) \\
&\qquad \times \mu_{\ell_i} e^{-\mu_{\ell_i}(m-\tau_0)} e^{-\mu_{\ell_{i+1}}(n-\tau_0)} \, dF(n) \, dm.
\end{aligned} \tag{19}$$

Because of the construction of policies π and π^1, we have for $m \leq n$

$$E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \sigma_i = n, \tau \geq n)$$
$$= E(V^{\pi^1} - V^\pi \,|\, \sigma_i = m, \sigma_{i+1} = n, \tau \geq n), \tag{20}$$
$$E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \tau = n, \sigma_i \geq n)$$
$$= E(V^{\pi^1} - V^\pi \,|\, \sigma_i = m, \tau = n, \sigma_{i+1} \geq n). \tag{21}$$

We also have for $m \leq n$

$$E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \sigma_i = n, \tau \geq n) \geq 0, \tag{22}$$

because only tardiness penalties are considered and the difference in the costs incurred under π and π^1 comes from the task that is completed at time m under π^1 and time n under π, that is, a task that is completed earlier under π^1. Finally, we have

$$E(V^\pi - V^{\pi^1} \,|\, \sigma_{i+1} = m, \tau = n, \sigma_i \geq n) \geq 0. \tag{23}$$

This is true for the following reason: from time m until time n, there is one extra task in the system under π (the one that completed service at time m under π^1). After time n, depending on the sequence of arrivals and service completions, there is either one extra task in the system under π (in service or waiting) or at some time the service of this task is completed and policies π and π^1 are coupled afterward. From Eqs. (19)-(23) and $\mu_{\ell_i} < \mu_{\ell_{i+1}}$, we get $E(V^{\pi^1}) \leq E(V^\pi)$.

We can now construct a modification π^2 of π^1 in the same way π^1 modifies π; that is, the first time π^1 prefers S_{ℓ_i} instead of $S_{\ell_{i+1}}$, π^2 uses $S_{\ell_{i+1}}$ instead of S_{ℓ_i} and follows the same priority rule as π^1 afterward. Therefore, $E(V^{\pi^2}) \leq E(V^{\pi^1})$. Continuing the construction of such modified policies, we conclude that $E(V^{\tilde{\pi}}) \leq E(V^\pi)$, where $\tilde{\pi}$ follows the priority rule $(S_{\ell_1}, \ldots, S_{\ell_{i-1}}, S_{\ell_{i+1}}, S_{\ell_i}, \ldots, S_{\ell_M})$.

If we keep improving policy $\tilde{\pi}$ by interchanging the order in each pair of consecutive servers for which the slowest server has the highest priority, we will eventually get that the optimal policy ranks the servers in decreasing order of their service rates. ■

If in Theorem 4 only earliness penalties are incurred, it is optimal to use the slowest servers, assigning the task with the earliest due date to the fastest among these servers. When both tardiness and earliness penalties are incurred, then, depending on the state of the system, it may be optimal to use either the fast or the slow servers. We illustrate this point by the following example.

Example 1: We have one task with due date d and two servers S_1 and S_2 with rates μ_1 and μ_2, respectively, where $\mu_1 > \mu_2$. Earliness and tardiness penalties are incurred at rates α and β, respectively. We have no discount ($\gamma = 0$) and no arrivals. Let π_1 and π_2 be the policies that assign d to servers S_1 and S_2, respectively. With x being the service completion time of task d, we get that the expected cost due to policy π_1 is

$$E(V^{\pi_1}) = \int_0^d \alpha(d-x)\mu_1 e^{-\mu_1 x}\,dx + \int_d^\infty \beta(x-d)\mu_1 e^{-\mu_1 x}\,dx$$

$$= \alpha d + \frac{\alpha}{\mu_1} e^{-\mu_1 d} - \frac{\alpha}{\mu_1} + \frac{\beta}{\mu_1} e^{-\mu_1 d}. \tag{24}$$

Similarly, the expected cost due to policy π_2 is

$$E(V^{\pi_2}) = \alpha d + \frac{\alpha}{\mu_2} e^{-\mu_2 d} - \frac{\alpha}{\mu_2} + \frac{\beta}{\mu_2} e^{-\mu_2 d}. \tag{25}$$

From Eqs. (24) and (25), we get for $d \to 0$

$$E(V^{\pi_1}) - E(V^{\pi_2}) \to \frac{\beta}{\mu_1} - \frac{\beta}{\mu_2} < 0$$

and for $d \to \infty$

$$E(V^{\pi_1}) - E(V^{\pi_2}) \to \frac{\alpha}{\mu_2} - \frac{\alpha}{\mu_1} > 0.$$

We see from this example that when d is sufficiently small it is optimal to use the fast server, and when it is sufficiently large it is optimal to use the slow server.

3.2. Stochastic Due Dates

The development of Section 3.2 parallels that of Section 3.1. We maintain the same assumptions as in Section 3.1, namely, that there are M parallel exponential servers $S_1, S_2, \ldots, S_M$, with rates $\mu_1, \mu_2, \ldots, \mu_M$, respectively.

We consider first a decision instant where the number of tasks is greater than or equal to the number of servers. Optimal policies are characterized by the following theorem.

THEOREM 5:

(i) *Let t_0 be a decision instant for which the number of tasks is greater than or equal to the number of servers, and let S_j be the fastest server. Then, within a class of tasks it is optimal to assign to S_j the one with the earliest arrival time.*

(ii) *In the single-server problem ($M = 1$), for each pair of classes i, j ($i, j = 1, 2, \ldots, N, i \neq j$) with $M^i(t_0) \neq \emptyset$, $M^j(t_0) \neq \emptyset$, there exists a time instant (threshold) t_{ij} such that it is optimal to process task a_1^i instead of a_1^j if $a_1^j \geq t_{ij}$, and vice versa otherwise.*

PROOF:

(i) Consider a class i of tasks with at least two tasks present in the system at time t_0. Let π be a policy that at t_0 assigns to S_j task $a_k^i, k \neq 1$, and proceeds optimally after t_0. We construct policy $\tilde{\pi}$ as follows: at t_0 $\tilde{\pi}$ assigns tasks in the same way as π, except that it assigns a_1^i to S_j and, if π has assigned a_1^i to some other server, it assigns a_k^i to the same server. After t_0 $\tilde{\pi}$ proceeds optimally. We also construct policy π' as follows: at t_0 π' assigns a_1^i to S_j and is identical to π when π processes tasks other than a_1^i. When π processes (if ever) a_1^i at some server, π' processes a_k^i at the same server.

We consider first the case when π assigns a_1^i to some server S_ℓ at t_0. Similar to the proof of Theorem 3, we get

$$E(V^\pi - V^{\tilde{\pi}})$$
$$= \int_{t_0}^{\infty} E(V_k^\sigma - V_1^\sigma) P(A^\sigma) \prod_m e^{-\mu_m(\sigma - t_0)} (\mu_j - \mu_\ell)\, d\sigma, \qquad \textbf{(26)}$$

where

$$A^\sigma = \{\omega \in \Omega \,|\, \text{first arrival after } t_0 \text{ occurs after time } \sigma\},$$

and V_k^σ (respectively, V_1^σ) is the cost incurred by the optimal policy, denoted by π_k^* (respectively, π_1^*), after time σ given that task a_k^i (respectively, a_1^i) completes its service at time σ and no other event has occurred until time σ. We construct policy $\tilde{\pi}_1^*$ as follows: $\tilde{\pi}_1^*$ is identical to π_k^* when π_k^* processes tasks other than a_1^i. When π_k^* processes (if ever) a_1^i, $\tilde{\pi}_1^*$ processes a_k^i. Let $V^{\tilde{\pi}_1^*}$ be the cost incurred under $\tilde{\pi}_1^*$ after σ. As in the proof of Theorem 3, with $a_1^i + x_1^i$ and $a_k^i + x_k^i$ being the due dates of a_1^i and a_k^i, respectively, we get

$$E(V_k^\sigma - V^{\tilde{\pi}_1^*}) = \frac{\alpha_i + \beta_i}{\gamma} E[e^{-\gamma \max\{a_1^i + x_1^i, \sigma\}} + e^{-\gamma \max\{a_k^i + x_k^i, \tau\}}$$
$$- e^{-\gamma \max\{a_k^i + x_k^i, \sigma\}} - e^{-\gamma \max\{a_1^i + x_1^i, \tau\}}].$$

The expression over which the expectation is taken is nonnegative for $a_1^i < a_k^i$ and $\sigma \leq \tau$. Therefore,

$$E(V_k^\sigma) \geq E(V^{\tilde{\pi}_1^*}) \geq E(V_1^\sigma), \qquad \textbf{(27)}$$

where the second inequality follows from the fact that policy π_1^* is optimal, while $\tilde{\pi}_1^*$ is not necessarily optimal. From Eqs. (26) and (27) and $\mu_j \geq \mu_\ell$, we conclude that $E(V^\pi) \geq E(V^{\tilde{\pi}})$.

For the case when π does not process a_1^i at t_0, using the same arguments as in the proof of part (i) of Theorem 3, we can show that $E(V^\pi) \geq E(V^{\tilde{\pi}})$.

Therefore, among all tasks of type i it is optimal to assign to S_j the one with the earliest arrival time.

(ii) Let a^i and a^j be tasks of type i and j that are present in the system at time t_0. Let the arrival times of all tasks except a^j be fixed. It suffices to show that if for $a^j = k$ it is optimal to process a^i instead of a^j, then for $a^j = \ell$, where $\ell > k$, it is still optimal to process a^i.

The proof uses the same arguments as the proof of part (ii) of Theorem 3, because for a specific sample path the due date of a^j when $a^j = k$ and $a^j = \ell$ is $k + x_j$ and $\ell + x_j$, respectively, where x_j is the realization of the random variable X_j for this sample path. ■

For decision instants where the number of tasks is smaller than the number of servers, we restrict attention to list scheduling strategies, that is, strategies that select the servers according to a prespecified priority rule $(S_{\ell_1}, S_{\ell_2}, \ldots, S_{\ell_M})$, where $(\ell_1, \ell_2, \ldots, \ell_M)$ is a permutation of $(1,2,\ldots,M)$. Properties of the optimal policy are described by the following theorem.

THEOREM 6: *Assume that we have only one class of tasks and only tardiness penalties are incurred. Let t_0 be a decision instant for which the number of tasks is smaller than the number of servers. Then*:

(i) *Once the servers have been selected, it is optimal to assign the task with the earliest arrival time to the fastest server.*

(ii) *The optimal list scheduling policy ranks the servers in decreasing order of their service rates, that is, uses the fastest servers.*

PROOF:

(i) The proof uses the same arguments as the proof of part (i) of Theorem 5.

(ii) The proof is identical to the proof of part (ii) of Theorem 4. ■

Theorems 4 and 6 exhibit the same similarities as Theorems 1 and 2; that is, one can get the statement of Theorem 6 from the statement of Theorem 4 by replacing due dates with arrival times.

If in Theorem 6 only earliness penalties are incurred, it is optimal to use the slowest servers, assigning the task with the earliest arrival time to the fastest among these servers. When both tardiness and earliness penalties are incurred, it may be optimal to use either the fast or the slow servers, as the following example illustrates.

Example 2: Consider decision instant t_0. We have one task with arrival time ρ and two servers S_1 and S_2 with rates $\mu_1 = 2$ and $\mu_2 = 1$, respectively. Earliness and tardiness penalties are incurred at rates $\alpha = 10, \beta = 1$. We have no discount $(\gamma = 0)$ and no arrivals. Due dates are exponential with rate 1. Let π_1 and π_2 be the policies that assign task ρ to S_1 and S_2, respectively. With d being the due date of ρ, we get for the expected cost of policy π_1

$$E(V^{\pi_1}) = E(V^{\pi_1} \mid d < t_0)P(d < t_0) + \int_{t_0}^{\infty} E(V^{\pi_1} \mid d = m)e^{-(m-\rho)}\,dm. \tag{28}$$

Let x be the service completion time of ρ. Then,

$$E(V^{\pi_1} \mid d < t_0) = \int_{t_0}^{\infty} \beta(x - t_0)\mu_1 e^{-\mu_1(x-t_0)}\,dx = \frac{\beta}{\mu_1}. \tag{29}$$

As in Example 1, we can show that

$$E(V^{\pi_1} \mid d = m, m > t_0) = \alpha(m - t_0) + \frac{\alpha + \beta}{\mu_1} e^{-\mu_1(m-t_0)} - \frac{\alpha}{\mu_1}. \tag{30}$$

From Eqs. (28)-(30), we get

$$E(V^{\pi_1}) = \frac{\beta}{\mu_1} - \frac{\alpha + \beta}{1 + \mu_1} e^{-(t_0-\rho)} + \alpha e^{-(t_0-\rho)}. \tag{31}$$

Similarly, the expected cost of policy π_2 is

$$E(V^{\pi_2}) = \frac{\beta}{\mu_2} - \frac{\alpha + \beta}{1 + \mu_2} e^{-(t_0-\rho)} + \alpha e^{-(t_0-\rho)}. \tag{32}$$

From Eqs. (31) and (32), we get for $\rho \to t_0$

$$E(V^{\pi_1}) - E(V^{\pi_2}) \to \frac{\beta}{\mu_1(1 + \mu_1)} - \frac{\beta}{\mu_2(1 + \mu_2)} - \frac{\alpha}{1 + \mu_1} + \frac{\alpha}{1 + \mu_2} > 0 \tag{33}$$

and for $\rho \to -\infty$

$$E(V^{\pi_1}) - E(V^{\pi_2}) \to \frac{\beta}{\mu_1} - \frac{\beta}{\mu_2} < 0. \tag{34}$$

We see from this example that when ρ is sufficiently small it is optimal to use the fast server, and when it is sufficiently large it is optimal to use the slow server.

4. CONCLUSIONS

A careful comparison of the results for deterministic due dates and stochastic due dates reveals many similarities between these two classes of problems. The role of due dates in deterministic due date problems is replaced by the arrival times in stochastic due date problems. The results of this paper referring to stochastic due dates can be obtained from the corresponding results for deterministic due dates by replacing due dates with arrival times.

The qualitative properties of optimal policies derived in this paper can guide the search for optimal dynamic scheduling strategies in multi-class queues with earliness and/or tardiness constraints. However, the determination of optimal scheduling policies still remains a formidable task, because, in general, optimal thresholds may depend on all the tasks present in the system as well as on the arrival process.

Acknowledgment

The authors thank the anonymous reviewer whose comments helped strengthen results of the paper.

References

1. Altman, E. & Shwartz, A. (1989). Optimal priority assignment: A time sharing approach. *IEEE Transactions on Automatic Control* 34: 1098–1102.
2. Baccelli, F., Boyer, P., & Hebuterne, G. (1984). Single server queues with impatient customers. *Advances in Applied Probability* 16: 887–905.
3. Baccelli, F. & Trivedi, K.S. (1985). A single server queue in a hard real time environment. *Operations Research Letters* 4: 161–168.
4. Baker, K.R. & Scudder, G.D. (1990). Sequencing with earliness and tardiness penalties: A review. *Operations Research* 38: 22–36.
5. Bhattacharya, P.P. & Ephremides, A. (1989). Optimal scheduling with strict deadlines. *IEEE Transactions on Automatic Control* 34: 721–728.
6. Bhattacharya, P.P. & Ephremides, A. (1991). Optimal allocation of a server between two queues with due times. *IEEE Transactions on Automatic Control* 36: 1417–1423.
7. Bhattacharya, P.P. & Ephremides, A. (1991). Stochastic monotonicity properties of multiserver queues with impatient customers. *Journal of Applied Probability* 28: 673–682.
8. Du, J. & Leung, J. (1990). Minimizing total tardiness on one machine is NP-hard. *Mathematics of Operations Research* 15: 483–495.
9. Frostig, E. (1991). A note on stochastic scheduling on a single machine subject to breakdown – The preemptive repeat model. *Probability in the Engineering and Informational Sciences* 5: 349–354.
10. Hall, N.G., Kubiak, W., & Sethi, S. (1991). Earliness–tardiness scheduling problems, II: Deviation of completion times about a restrictive common due date. *Operations Research* 39: 847–856.
11. Hall, N.G. & Posner, M.E. (1991). Earliness–tardiness scheduling problems, I: Weighted deviation of completion times about a common due date. *Operations Research* 39: 836–846.
12. Huang, C.-C. & Weiss, G. (1992). Scheduling jobs with stochastic processing times and due dates to minimize total tardiness. Preprint.
13. Kaspi, H. & Perry, D. (1983). Inventory systems of perishable commodities. *Advances in Applied Probability* 15: 674–685.
14. Kaspi, H. & Perry, D. (1983). Inventory systems of perishable commodities with renewal input and Poisson output. *Advances in Applied Probability* 16: 402–421.
15. Kubiak, W., Lou, S., & Sethi, S. (1990). Equivalence of mean flow time problems and mean absolute deviation problems. *Operations Research Letters* 9: 371–374.
16. Nain, P. & Ross, K.W. (1986). Optimal priority assignment with hard constraint. *IEEE Transactions on Automatic Control* 31: 883–888.
17. Pandelis, D. (1994). Optimal stochastic scheduling and routing in queueing networks. Doctoral dissertation, University of Michigan, Ann Arbor.
18. Pandelis, D.G. & Teneketzis, D. (1993). Stochastic scheduling in priority queues with strict deadlines. *Probability in the Engineering and Informational Sciences* 7: 273–289.
19. Panwar, S.S., Towsley, D., & Wolff, J.K. (1988). Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service. *Journal of Association for Computing Machinery* 35: 832–844.
20. Perry, D. & Levikson, B. (1989). Continuous production/inventory model with analogy to certain queueing and dam models. *Advances in Applied Probability* 21: 123–141.

21. Pinedo, M. (1983). Stochastic scheduling with release dates and due dates. *Operations Research* 31: 559–572.
22. Potts, C.N. & Van Wassenhove, L.N. (1992). Single machine scheduling to minimize total late work. *Operations Research* 40: 586–595.
23. Ross, K.W. & Chen, B. (1988). Optimal scheduling of interactive and noninteractive traffic in telecommunication sytems. *IEEE Transactions on Automatic Control* 33: 261–267.
24. Szwarc, F.W. (1990). Parametric precedence relations in single machine scheduling. *Operations Research Letters* 9: 133–140.
25. Towsley, D. & Baccelli, F. (1991). Comparisons of service disciplines in a tandem queueing network with real time constraints. *Operations Research Letters* 10: 49–55.