Optimal Assignment of Impatient Customers to Parallel Queues with Blocking

A. Movaghar[1]

In this paper, a classical routing problem in a time-critical environment is studied. A state-dependent queueing system with a single arrival stream and a number of identical servers is considered. Each server has its own queue with a finite capacity. No jockeying among queues is allowed. Upon arrival, a customer must join a server's queue; if all queues are full, it will leave the system immediately and is considered lost. Each arriving customer is limited to an exponentially distributed patience time after which it must depart the system and becomes lost. Two models of customer behavior are considered: deadlines until the beginning of service and deadlines until the end of service. An arriving customer is assigned to a server's queue, in order to minimize the number of lost customers. It is proven that the policy of joining shortest non-full queue (SNQ) minimizes, with respect to stochastic order, the number of lost customers by any time. For the second model of customer behavior, a stronger result, namely, that SNQ policy minimizes, with respect to stochastic order, the number of customers which miss their deadlines as well as the number of customers which are blocked due to full queues by any time is also established.

INTRODUCTION

A classical routing problem is how to optimally route a stream of customers among a number of parallel queues. Consider a queueing system with a Poisson arrival process and a finite number of identical exponential servers. Each server has its own queue. No jockeying among queues is allowed. Upon arrival, a customer must join a server's queue. An interesting problem is how to optimally assign an arriving customer to a server's queue. This problem was first studied by Winston [1] who showed that the policy of joining shortest queue (SQ) maximizes, with respect to stochastic order, the discounted number of customers served by any time. Davis [2] generalized this result to renewal arrival process. Weber [3], subsequently, extended this latter result to an arbitrary arrival process and a general service-time distribution with an increasing hazard rate. Ephremides et al. [4] showed that, for a system with an arbitrary arrival process and two identical exponential servers, SQ policy minimizes the expected system delay by any time. Walrand [5] extended this result to a finite number of servers, using a simple forward induction method. He also showed that SQ policy stochastically minimizes the number of customers in the system by any time. Johri [6] established similar results for a model with Poisson arrival process and state-dependent service rates, where each server has a rate which is a bounded, increasing and concave function

1. Department of Computer Engineering, Sharif University of Technology, Tehran, I.R. Iran.

Scientia Iranica, Vol. 3, No. 4, © Sharif University of Technology, January 1997.

of the number of customers in its queue. Whitt [7] gave some counterexamples proving that SQ policy is not necessarily optimal for all general service-time distributions to minimize a customer's expected delay or the expected number of customers served in a given time. The optimality of SQ policy has also been established by Menich and Serfozo [8] for a Markovian model with state-dependent arrival process, service rates and holding cost rate.

Hordijk and Koole [9] have more recently considered a finite capacity model with a state-dependent arrival process and identical exponential servers. They proved that the policy of joining shortest non-full queue (SNQ), a generalization of SQ policy, stochastically maximizes the number of customers served by any time. Towsley et al. [10] considered a similar model and showed that SNQ policy stochastically minimizes the number of customers which are blocked due to full queues by any time. Sparaggis et al. [11] extended this latter result to models with state-dependent service-rates, where each server has a rate which is an increasing and concave function of the number of customers in its queue.

All of the above studies, however, share a common implicit assumption, namely, that the environment of the system is not time-critical. In such environments, customers are "patient", i.e., they stay in the system until they are served. On the other hand, with the advent of many important real-time applications (such as computer communication networks, industrial process control, automated manufacturing, avionics, traffic control, etc.), there has been a growing interest in environments which are time-critical. In such environments, customers are usually "impatient", i.e., they may not stay in the system indefinitely. Each impatient customer has a deadline after which it must leave the system and is considered lost. Queues with impatient customers have been moderately analyzed [12–17]. Recently, there have also been some results about the optimal control of such queues [18,19]. The principal objective in the latter works is to minimize the number of lost customers.

In this paper, a time-critical environment is assumed. A state-dependent queueing system is considered with a single arrival stream and s number of identical servers. Each server has its own queue. The capacity of i-th server's queue is M_i ($M_i < \infty$), $i = 1, \ldots, s$. The service discipline in each queue is first-come-first-served (FCFS). No jockeying among queues is allowed. Let the number of customers in i-th server's queue be n_i, $n_i = 1, \ldots, M_i$, $i = 1, \ldots, s$. Then, the rate of i-th server is $\mu(n_i)$, $i = 1, \ldots, s$, where μ is an increasing and concave function. The arrival rate is $\lambda(n_1, \ldots, n_s)$, where λ is a decreasing and Schur-convex function. Upon arrival, a customer must join a server's queue; if all queues are full, it will leave the system immediately and is considered lost. Each arriving customer has a deadline after which it must leave the system and becomes lost. The difference between the deadline of a customer and its arrival time, referred to as a "relative deadline", is an exponentially distributed random variable with rate ν. Relative deadlines are assumed to be i.i.d. random variables and independent of the arrival process and service times. Two models of customer behavior are considered: deadlines until the beginning of service and deadlines until the end of service. In the first model, a customer keeps its deadline only until the beginning of its service. Accordingly, customers remain in the system while being served until they complete their service requirements. In the second model, a customer retains its deadline until the end of its service. Accordingly, customers may discontinue their services because they have missed their deadlines. Assigning an arriving customer to a server's queue is sought so as to minimize the number of lost customers. It is proven that SNQ policy minimizes, with repect to stochastic order, the number of lost customers by any time. For the second model of customer behavior, a stronger result is also established, namely, that SNQ policy minimizes, with respect to stochastic order, the number of customers which miss their deadlines as well as the number of customers which are blocked due to full queues by any time.

This paper is organized as follows. In the following section, the underlying mathematical models used in this paper are described. Then, the optimality of SNQ policy is proven. Finally, the main results of the paper are summarized.

MODELS

In this section, the modeling framework is presented and certain aspects of some mathematical concepts are outlined. This framework is closely related to the theory of Markov decision processes [20,21], but differs from the latter in that it considers the state-transitions of the model more closely, due to activity completions. Throughout this paper, R denotes the set of real numbers and R_+ represents the set of non-negative real numbers.

Definition 1

A (continuous-time) controlled Markov system is a structure $\mathcal{M} = (Q, \Delta, A, r)$ where

- Q is a countable set of states,
- Δ is a countable set of control actions,
- A is a finite set of activities,
- $r = \{r(.,.|q,\delta);\ q \in Q, \delta \in \Delta\}$ such that for any $q \in Q$ and $\delta \in \Delta$,

$$r(.,.|q,\delta):\ A \times Q \longrightarrow R_+.$$

The above model represents a dynamic system which behaves in time as follows. Suppose, $\mathcal{M}$ is observed during $[0,\infty)$. At any time, $\mathcal{M}$ is in some state and under some control action. At any time, at most, one activity may be completed. A state or control action changes only if an activity completes. For $t, \epsilon \in R_+$, $q, q' \in Q$, $\delta \in \Delta$ and $a \in A$, let:

$\xi_\epsilon^t(a, q'|q,\delta) \equiv$ the probability that activity a completes during $[t, t+\epsilon)$ and the resulting state is q', given that the state and control action at time t are q and δ, respectively,

$O_\epsilon^t(q,\delta) \equiv$ the probability that more than one activity complete during $[t, t+\epsilon)$, given that the state and control action at time t are q and δ, respectively.

Then, the state changes are governed by the following properties:

$$\lim_{\epsilon \to 0} \frac{\xi_\epsilon^t(a,q'|q,\delta)}{\epsilon} = r(a, q'|q,\delta)\ ,$$

$$\lim_{\epsilon \to 0} \frac{O_\epsilon^t(q,\delta)}{\epsilon} = 0\ .$$

The control actions are determined as follows. Let X_t, Y_t and Z_t represent the state, control action and activity completed (if any) at time t, respectively. ($Z_t = \emptyset$ if no activity completes at time t.) It is noted that $H_t = \{X_{t'}, Y_{t'}, Z_{t'}; 0 \le t' < t, t' \in R_+\}$. H_t is called the history of $\mathcal{M}$ prior to t. A policy for $\mathcal{M}$ is a set $\Pi = \{\Pi_t; t \in [0,\infty)\}$ where, for any time t, $\Pi_t(.|X_t, H_t)$ is a conditional probability measure over Δ. Suppose, activity a completes immediately before time t. Then, at time t, policy Π chooses a control action δ_t with probability $\Pi_t(\delta_t|X_t, H_t)$. Π is said to be a Markov policy if, for any time t, $\Pi_t(.|X_t, H_t)$ is independent of H_t and is a stationary policy if $\Pi_t(.|X_t, H_t)$ is independent of t. Π is called a non-randomized or deterministic policy if, for any time t, $\Pi_t(\delta_t|X_t, H_t) = 1$ for some control action δ_t. Given X_0 and a policy Π, $\Psi = \{X_t, Y_t, Z_t; t \in \mathcal{R}_+\}$ is a well-defined random process, called a controlled Markov process. Ψ is assumed to be "regular" in the sense that during any finite interval, with probability 1, at most a finite number of activity completions may occur. The existence of a cost structure $c = (c', c'')$ is postulated, where c' is a bounded Borel measurable function,

$$c' : Q \times \Delta \longrightarrow R_+,$$

referred to as a cost rate function and c'' is a bounded Borel measurable function,

$$c'' : Q \times \Delta \times (A \cup \emptyset) \times Q \longrightarrow R_+,$$

called as an activity cost function. Let D_t and t^+ denote the number of activity completions by time t and a time immediately after t, respectively. The cost incurred will be of interest during a period $[\tau', \tau]$, $\tau', \tau \in R_+$, $\tau' \le \tau$, defined as:

$$C = \int_{\tau'}^{\tau} c'(X_t, Y_t)dt + \int_{\tau'}^{\tau} c''(X_t, Y_t, Z_t, X_{t^+})\, dD_t, \quad (1)$$

where C is a well-defined random variable. Accordingly, to compare various costs, a notion of ordering among random variables is needed.

Definition 2

Let X and Y be two real-valued random variables. X is said to be stochastically greater than (or equal to) Y, denoted as $X \geq_{st} Y$, if:

$$P[X > x] \geq P[Y > x],$$

for all $x \in R$.

Let $(\mathcal{M}, c)$ denote a controlled Markov system $\mathcal{M}$ with a cost structure c. For any $t \in R_+$, a policy Π and a state q define:

$$C^t(\Pi|q) \equiv \text{the cost incurred during } [0,t], \text{ given an initial state } q. \quad (2)$$

A policy Π^* is sought such that for any policy Π and state q,

$$C^t(\Pi|q) \geq_{st} C^t(\Pi^*|q). \quad (3)$$

Π^* is called an optimal policy for $(\mathcal{M}, c)$ by time t. Such optimal policy, however, may not always exist.

This section is concluded with a brief outline of some mathematical concepts which are used in this paper. Let $N = \{0, 1, \dots\}$ represent the set of natural numbers and $N^K = \{(x_1, \dots, x_K);\ x_i \in N,\ i = 1, \dots, K\}$ the set of all K-tuples of natural numbers.

Definition 3

A function $\phi : N \to R$ is said to be convex if for every $\alpha, \overline{\alpha} \in [0,1]$ and $x, y, z \in N$ such that $\alpha + \overline{\alpha} = 1$ and $z = \alpha x + \overline{\alpha} y$,

$$\alpha \phi(x) + \overline{\alpha} \phi(y) \geq \phi(z).$$

ϕ is said to be a concave function if $(-\phi)$ is convex. ϕ is called an affine function if it is both convex and concave.

Lemma 1

A function $\phi : N \to R$ is convex iff, for any $x, y, \overline{x}, \overline{y} \in N$ such that $\overline{x} \leq x \leq y \leq \overline{y}$ and $x + y = \overline{x} + \overline{y}$,

$$\phi(\overline{x}) + \phi(\overline{y}) \geq \phi(x) + \phi(y). \quad (4)$$

Next, the notion of Schur-convex functions is considered. (For a more complete treatment of this concept, please see [22].) For any $x = (x_1, \dots, x_K) \in N^K$, let,

$$x_{[1]} \geq \dots \geq x_{[K]},$$

denote the components of x in a decreasing order.

Definition 4

For any $x, y \in N^K$, it is said that y majorizes x (or x is majorized by y), denoted by $x \preceq y$, if:

$$\sum_{i=1}^{k} x_{[i]} \leq \sum_{i=1}^{k} y_{[i]}, k = 1, \dots, K-1,$$

$$\sum_{i=1}^{K} x_{[i]} = \sum_{i=1}^{K} y_{[i]}.$$

Definition 5

A function $\phi : N^K \to R$ is said to be a Schur-convex function if, for any $x, y \in N^K$ such that $x \preceq y$,

$$\phi(x) \leq \phi(y).$$

Lemma 2

Let $\phi : N^K \to R$ be a Schur-convex function. For $x = (x_1, \dots, x_K) \in N^K$, define $\sigma x = (x_{\sigma(1)}, \dots, x_{\sigma(K)})$, where σ is a permutation of integers $1, \dots, K$. Then,

$$\phi(x) = \phi(\sigma x). \quad (5)$$

Finally, the notion of monotone functions is reviewed. For $x, y \in N^K$, $x = (x_1, \dots, x_K)$, $y = (y_1, \dots, y_K)$, let $x \leq y$ denote the elementwise ordering, i.e., $x_i \leq y_i$, for all i, $i = 1, \dots, K$.

Definition 6

A function $\phi : N^K \to R$ is said to be increasing if for any $x, y \in N^K$ such that $x \leq y$,

$$\phi(x) \leq \phi(y).$$

ϕ is called decreasing if $(-\phi)$ is increasing.

OPTIMAL POLICY

Consider the queueing problem discussed previously. Customers become lost either because they miss their deadlines or they are blocked due to full queues. In this section, it is proven that the policy of joining shortest non-full queue (SNQ) minimizes, with repect to stochastic order, the number of lost customers by any time. For the second model of customer behavior, a stronger result is also establish, namely, that SNQ policy stochastically minimizes the number of customers which miss their deadlines, as well as the number of customers which are blocked due to full queues by any time. To show these, the general modeling framework described before is employed. Also an additional "dummy" activity is introduced. This extra activity permits a new formulation of the problem which is more amenable to analysis. Use of dummy activities in optimal control of queues is not new and has also been reported in [5,23].

Let $q = (n_1, \ldots, n_s)$ be a s-tuple of natural numbers. The following notations are used:

- $q_i = n_i$, $i = 1, \ldots, s$,
- $q + i \equiv (n_1, \ldots, n_i + 1, \ldots, n_s)$, where $n_i < M_i$, $i = 1, \ldots, s$,
- $q - i \equiv (n_1, \ldots, n_i - 1, \ldots, n_s)$, where $n_i > 0$, $i = 1, \ldots, s$.

The queueing problem described in the previous section may now be modeled as follows. Let $\mathcal{M} = (Q, \Delta, A, r)$ be a controlled Markov system such that:

- $Q = \{(n_1, \ldots, n_s); n_i = 0, \ldots, M_i, i = 1, \ldots, s\}$, where a state $q = (n_1, \ldots, n_s)$ means that there are n_i customers in i-th server's queue, $i = 1, \ldots s$. $\emptyset = (0, \ldots, 0)$ denotes the empty state and $M = (M_1, \ldots, M_s)$ the full state,
- $A = \{a_0, a_1, a_2^i, a_3^i; i = 1, \ldots, s\}$ where:
  - $a_0 \equiv$ the dummy activity,
  - $a_1 \equiv$ the arrival of a customer,
  - $a_2^i \equiv$ the completion of a service by i-th server, $i = 1, \ldots, s$,
  - $a_3^i \equiv$ the rejection of a late customer in i-th server's queue, $i = 1, \ldots, s$,
- $\Delta = \{1, 2, \ldots, s\}$, where control action i means that an arriving customer is assigned to i-th server's queue, $i = 1, \ldots, s$,
- For any $q, q' \in Q$, $\delta \in \Delta$ and $i = 1, \ldots, s$,

$$r(a_0, q' | q, \delta) = \begin{cases} \alpha(q), & \text{where } q' = q, \\ 0, & \text{otherwise,} \end{cases}$$

$$r(a_1, q' | q, \delta) = \begin{cases} \lambda(q), & \text{where } q' = q + \delta \\ & \text{and } q_\delta < M_\delta, \\ \lambda(q), & \text{where } q' = q \\ & \text{and } q_\delta = M_\delta, \\ 0, & \text{otherwise,} \end{cases}$$

$$r(a_2^i, q' | q, \delta) = \begin{cases} \mu(q_i), & \text{where } q' = q - i \\ & \text{and } q_i > 0, \\ 0, & \text{otherwise,} \end{cases}$$

$$r(a_3^i, q' | q, \delta) = \begin{cases} \gamma(q_i), & \text{where } q' = q - i \\ & \text{and } q_i > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $\gamma(q_i)$ is the conditional rate of a customer missing its deadline in i-th server's queue, given that this queue has q_i customers and $\alpha(q)$ is the rate of the dummy activity in state q as defined below. Let ϵ be an arbitrary small positive real number. Then, $\gamma(n)\,\epsilon$ represents the probability that a customer in a server's queue misses its deadline during an interval of length ϵ, given that there are n customers in that queue. Since customer relative deadlines are exponentially distributed, and exponential distribution functions are memoryless, the following is obtained:

$$\gamma(n) = \nu(n-1), n = 1, 2, \ldots,$$
$$\gamma(0) = 0, \tag{6}$$

for the first model of customer behavior and:

$$\gamma(n) = \nu n, n = 0, 1, \ldots, \tag{7}$$

for the second model of customer behavior. (See [12,13] for more clarification.) $\alpha(q)$ is

defined as:

$$\alpha(q) = \lambda(\emptyset) - \lambda(q) + \sum_{i=1}^{s}[\mu(M_i) + \gamma(M_i) - \mu(q_i) - \gamma(q_i)] .$$

Note that $\alpha(q)$, the rate of the dummy activity, is chosen so that the overall activity rate of the system remains constant in all states.

Define cost structure $c = (c', c'')$ such that for any $q, q' \in Q$, $\delta \in \Delta$, $a \in A$, and $i = 1, \ldots, s$,

$$c' = 0 ,$$

$$c''(q, \delta, a, q') = \begin{cases} 1, & \text{where } a \in \{a_1^i, a_3^i\}, \\ & q' = q - i \text{ and } q_i > 0 , \\ 0, & \text{otherwise} . \end{cases}$$

$(\mathcal{M}, c)$ represents the problem of minimizing the number of lost customers. It is proven that SNQ policy is optimal for this problem. More formally, for any $q \in Q$, $q \neq M$, let:

$$W(q) = \min\{q_i;\ q_i < M_i, i = 1, \ldots, s\},$$

$$L(q) = \min\{i;\ q_i = W(q), q_i < M_i, i = 1, \ldots, s\}.$$

$W(q)$ represents the number of customers in SNQ and $L(q)$ the index of the first such queue. Define a stationary policy

$$\Pi'' = \{\Pi''_t; t \in R_+\} ,$$

such that, for any $q \in Q$ and $\delta \in \Delta$, $\Pi''_0(\delta|q) = 1$ iff either $q \neq M$ and $\delta = L(q)$, or $q = M$ and $\delta = 1$. Π'' represents a policy of joining shortest non-full queue. It shall be proven that Π'' is an optimal policy for $(\mathcal{M}, c)$ by any time t. To show this, the set of activity completion times by t is conditioned. For an arbitrary set of such activity completion times T, the optimality of Π'' is then demonstrated. It is also noted that the overall activity rate (i.e., the sum of all activity rates) in each state is fixed and is equal to:

$$\eta = \lambda(\emptyset) + \sum_{i=1}^{s}[\mu(M_i) + \gamma(M_i)], \tag{8}$$

which means that the overall activity completions constitute a Poisson process with rate η. Since T is arbitrary and η is fixed, Π'' can finally be shown to be optimal. Before embarking to prove the optimality of Π'', some helpful notations are introduced. In the sequel, it is assumed that Π is a policy, $t \in R_+$, T is a set of activity completion times by time t, $m \in N$, $m \leq |T|$, $q \in Q$ and h is a history. Let t_m represent the time of the m-th activity completion. ($t_0 = 0$.) Denote $C_m^t(\Pi|q, h; T)$ to be the cost incurred during $[t_m, t]$ for $(\mathcal{M}, c)$, given that q is the state at time t_m, h is the history prior to t_m and T is the set of activity completion times by time t. When Π is stationary, the above notation can be written more simply as $C_m^t(\Pi|q; T)$. For $q \in Q$, $q \neq M$, also the following notation is adopted:

$$q^* = q + L(q),$$

where q^* represents the state of the system under policy Π'', after a customer arrives in a (non-full) state q. Finally, let:

$$\overline{\gamma} = \gamma/\eta, \qquad \overline{\lambda} = \lambda/\eta,$$

$$\overline{\mu} = \mu/\eta, \qquad \overline{\alpha} = \alpha/\eta,$$

where η is defined as in Equation 8.

Lemma 3

$$C_m^t(\Pi|q, h; T) = \begin{cases} C_{m+1}^t(\Pi|q - i, h_2^i; T), \\ \quad \text{with prob. } \overline{\mu}(q_i) \text{ where} \\ \quad q_i > 0, i = 1, \ldots, s, \\ 1 + C_{m+1}^t(\Pi|q - i, h_3^i; T), \\ \quad \text{with prob. } \overline{\gamma}(q_i) \text{ where} \\ \quad q_i > 0, i = 1, \ldots, s, \\ C_{m+1}^t(\Pi|q + \delta, h_1; T), \\ \quad \text{with prob. } \overline{\lambda}(q)\Pi_{t_m}(\delta|q, h') \\ \quad \text{where } q_\delta < M_\delta \text{ and } \delta \in \Delta, \\ 1 + C_{m+1}^t(\Pi|q, h_1; T), \\ \quad \text{with prob. } \overline{\lambda}(q)\Pi_{t_m}(\delta|q, h') \\ \quad \text{where } q_\delta = M_\delta \text{ and } \delta \in \Delta, \\ C_{m+1}^t(\Pi|q, h_0; T), \\ \quad \text{with prob. } \overline{\alpha}(q), \end{cases} \tag{9}$$

where h_0, h_1, h', h_2^i and h_3^i, $i = 1, \ldots, s$, are some appropriate histories.

Proof

The proof follows immediately from definitions of $(\mathcal{M}, c)$ and Π. □

Lemma 4

$$C_m^t(\Pi''|q;T) = \begin{cases} C_{m+1}^t(\Pi''|q-i;T), \\ \quad \text{with prob. } \overline{\mu}(q_i) \text{ where} \\ \quad q_i > 0,\ i = 1, \ldots, s, \\ 1 + C_{m+1}^t(\Pi''|q-i;T), \\ \quad \text{with prob. } \overline{\gamma}(q_i) \text{ where} \\ \quad q_i > 0,\ i = 1, \ldots, s, \\ C_{m+1}^t(\Pi''|q^*;T), \\ \quad \text{with prob. } \overline{\lambda}(q) \text{ where} \\ \quad q \neq M, \\ 1 + C_{m+1}^t(\Pi''|q;T), \\ \quad \text{with prob. } \overline{\lambda}(q) \text{ where} \\ \quad q = M, \\ C_{m+1}^t(\Pi''|q;T), \\ \quad \text{with prob. } \overline{\alpha}(q), \end{cases} \tag{10}$$

Proof

Substitute Π, in the previous lemma, by Π''. The proof follows immediately from the definition of Π''. □

Lemma 5

Let q and $\overline{q}$ be two states such that $\overline{q}_i = q_{\sigma(i)}$, $i = 1, \ldots, s$, where σ is a permutation of integers $1, \ldots, s$. Then,

$$C_m^t(\Pi''|q;T) = C_m^t(\Pi''|\overline{q};T). \tag{11}$$

Proof

Induction on m is used. The lemma is true for $m = |T|$. It suffices to show that the lemma is true for m, given it is valid for $m+1$. Recall that λ is a Schur-convex function, which means $\overline{\lambda}$ is also a Schur-convex function. Thus, by Lemma 2, it is found that:

$$\overline{\lambda}(q) = \overline{\lambda}(\sigma q). \tag{12}$$

It is also noted that:

$$\sigma q - \sigma i = \sigma(q - i), \tag{13}$$

where $q_i > 0$, $i = 1, \ldots, s$ and that:

$$(\sigma q)^* = \sigma' q^*, \tag{14}$$

where σ' is a permutation of integers $1, \ldots, s$. Using Equations 10 and 12–14, the induction step is established, which proves the lemma. □

Lemma 6

Let q and $\overline{q}$ be two states such that $\overline{q_j} \leq q_j \leq q_l \leq \overline{q_l}$, $q_j + q_l = \overline{q_j} + \overline{q_l}$, for some j, l, $j \neq l$, $j = 1, \ldots, s$, $l = 1, \ldots, s$, $q_i = \overline{q_i}$, for all i, $i \neq j$, $i \neq l$, $i = 1, \ldots, s$, and some k, $k = 1, \ldots, s$, such that $q_k < M_k$. Then,

$$1 + C_m^t(\Pi''|q;T) \geq_{st} C_m^t(\Pi''|q+k;T), \tag{15}$$

$$C_m^t(\Pi''|q+k;T) \geq_{st} C_m^t(\Pi''|q;T), \tag{16}$$

$$C_m^t(\Pi''|\overline{q};T) \geq_{st} C_m^t(\Pi''|q;T). \tag{17}$$

Proof

Induction on m is used. The lemma is true for $m = |T|$. It is necessary to show that the lemma is true for m, given it is valid for $m + 1$.

Proof of induction step for Statements 15 and 16

Since μ and γ are increasing functions and λ is a decreasing function, $\overline{\mu}$ and $\overline{\gamma}$ are increasing and $\overline{\lambda}$ is decreasing. It is also noted that:

$$(q + k)^* = q^* + k', \tag{18}$$

for some k', $k' = 1, \ldots, s$. The proof follows from Equation 10 using the above properties.

Proof of induction step for Statement 17

When $\overline{q}_1 = q_1$, the proof is trivial. Thus, it is assumed that $\overline{q}_1 < q_1$. It is noted that μ is concave ($\mu(0) = 0$) and λ is Schur-convex. Furthermore, from Equations 6 and 7, it is found that γ is convex ($\gamma(0) = 0$). Accordingly, $\overline{\mu}$, $\overline{\gamma}$ and $\overline{\lambda}$ are also concave ($\overline{\mu}(0) = 0$), convex

($\overline{\gamma}(0) = 0$) and Schur-convex, respectively. Using Definition 5 and Lemma 1,

$$\overline{\mu}(\overline{q}_j) + \overline{\mu}(\overline{q}_l) \leq \overline{\mu}(q_j) + \overline{\mu}(q_l), \tag{19}$$

$$\overline{\gamma}(\overline{q}_j) + \overline{\gamma}(\overline{q}_l) \geq \overline{\gamma}(q_j) + \overline{\gamma}(q_l), \tag{20}$$

$$\overline{\lambda}(\overline{q}) \geq \overline{\lambda}(q). \tag{21}$$

Now, the following cases are considered:

1) $q_j = q_l$, $\overline{q_j} = 0$, 2) $q_j = q_l$, $\overline{q_j} > 0$,

3) $q_j < q_l$, $\overline{q_j} = 0$, 4) $q_j < q_l$, $\overline{q_j} > 0$.

From induction assumption, for Cases 1 and 3,

$$C^t_{m+1}(\Pi''|\overline{q} - l; T) \geq_{st} C^t_{m+1}(\Pi''|q - j; T) \geq_{st} C^t_{m+1}(\Pi''|q - l; T), \tag{22}$$

and for Cases 2 and 4,

$$C^t_{m+1}(\Pi''|\overline{q} - l; T) \geq_{st} C^t_{m+1}(\Pi''|\overline{q} - j; T) \geq_{st} C^t_{m+1}(\Pi''|q - j; T) \geq_{st} C^t_{m+1}(\Pi''|q - l; T). \tag{23}$$

Using Statements 10, 11 and 18–23, it suffices to prove that:

$$C^t_{m+1}(\Pi''|\overline{q}^*; T) \geq_{st} C^t_{m+1}(\Pi''|q^*; T). \tag{24}$$

For Cases 1 and 3, the following subcases are examined:

a) $L(\overline{q}) = L(q) = j$,

b) $L(\overline{q}) = j$, $L(q) = k$, $k = 1, \ldots, s$, $k \neq j$, $k \neq l$.

For Cases 2 and 4, the following subcases are considered:

c) $L(\overline{q}) = L(q) = j$,

d) $L(\overline{q}) = j$, $L(q) = k$, $k = 1, \ldots, s$, $k \neq j$, $k \neq l$,

e) $L(\overline{q}) = L(q) = k$, $k = 1, \ldots, s$, $k \neq j, k \neq l$.

From induction assumption and Equation 11, Statement 24 can be shown to be valid for each of the above subcases, which completes the proof. □

Theorem 1

Π'' is an optimal policy for $(\mathcal{M}, c)$ by any time t.

Proof

It must be shown that:

$$C^t(\Pi|q) \geq_{st} C^t(\Pi''|q), \tag{25}$$

for any policy Π and state q, where $C^t(\Pi|q)$ is defined as in Statement 2. It is recalled that the overall activity rate is fixed as in Equation 8. Accordingly, it is only required to prove that:

$$C^t_m(\Pi|q, h; T) \geq_{st} C^t_m(\Pi''|q; T), \tag{26}$$

where h is a history, T is a set of activity completion times by t and m is a natural number, $m \leq |T|$. To prove Statement 26, induction on m is used. This statement is true for $m = |T|$. To establish the induction step, it is necessary to show that Statement 26 is true for m, given it is valid for $m + 1$. Two cases are considered:

1) $q = M$,

2) $q < M$.

For case 1, the proof follows immediately from Equations 9 and 10. For case 2, let $q_k < M_k$, for some k, $k = 1, \ldots, s$. Using Equations 9 and 10, it is sufficient to prove that:

$$C^t_{m+1}(\Pi|q + k, h; T) \geq_{st} C^t_{m+1}(\Pi''|q^*; T). \tag{27}$$

From induction assumption, the following is obtained:

$$C^t_{m+1}(\Pi|q + k, h; T) \geq_{st} C^t_{m+1}(\Pi''|q + k; T). \tag{28}$$

Therefore, it suffices to establish that:

$$C^t_{m+1}(\Pi''|q + k; T) \geq_{st} C^t_{m+1}(\Pi''|q^*; T). \tag{29}$$

However, Statement 29 follows Statements 11 and 17, which concludes the proof. □

The above theorem proves that SNQ policy minimizes the number of lost customers by any

time. For the case of deadlines until the end of service, however, a stronger result can also be established, namely, that SNQ policy stochastically minimizes the number of customers which miss their deadlines, as well as the number of customers which are blocked due to full queues by any time. More formally, let cost structures $d = (d', d'')$ and $b = (b', b'')$ be defined such that for any $q, q' \in Q$, $\delta \in \Delta$, $a \in A$, and $i = 1, \ldots, s$,

$$d' = b' = 0 \ ,$$

$$d''(q, \delta, a, q') = \begin{cases} 1, & \text{where } a = a_3^i, \\ & q' = q - i \text{ and } q_i > 0 \ , \\ 0, & \text{otherwise,} \end{cases}$$

$$b''(q, \delta, a, q') = \begin{cases} 1, & \text{where } a = a_1, \ q' = q \\ & \text{and } q_\delta = M_\delta, \\ 0, & \text{otherwise.} \end{cases}$$

$(\mathcal{M}, d)$ and $(\mathcal{M}, b)$ represent the problem of minimizing the number of customers which miss their deadlines and the problem of minimizing the number of customers which are blocked due to full queues, respectively.

Theorem 2

For the second model of customer behavior, i.e., the case of customer deadlines until the end of service, Π'' is an optimal policy for $(\mathcal{M}, b)$ and $(\mathcal{M}, d)$ by any time t.

Proof

It is noted that γ, as defined in Equation 7, is an affine function. The proof is then similar to that of Theorem 1. □

CONCLUSIONS

The problem of optimally assigning a stream of impatient customers to parallel queues with finite capacities has been studied. Customers may become lost due to their impatience or the full queues. For exponentially distributed customer impatience, it has been shown that the policy of joining the shortest non-full queue is optimal in minimizing the number of lost customers. This is an interesting result regarding the problem of optimal routing in a real-time environment. However, it is still an open question whether similar results will hold for other types of customer impatience and work is currently underway in answer to this problem.

REFERENCES

1. Winston, W. "Optimality of the shortest line discipline", *J. of Appl. Prob.*, **14**, pp. 181-189 (1977).
2. Davis, E. "Optimal joining policies for a queueing system having parallel channels", Ph.D. Dissertation, North Carolina State University, USA (1977).
3. Weber, R. "On the optimal assignment of customers to parallel servers", *J. of Appl. Prob.*, **15**, pp 406-413 (1978).
4. Ephremides, A., Varaiya, P. and Walrand, J. "A simple dynamic routing problem", *IEEE Trans. Automat. Contr.*, **4**, pp 690-693 (1980).
5. Walrand, J. *An Introduction to Queueing Networks,* Prentice Hall, Inc., Englewood Cliffs, NJ, USA (1988).
6. Johri, P.K. "Optimality of the shortest line discipline with state-dependent service times", *European J. of Oper. Res.*, **41**, pp 157-161 (1989).
7. Whitt, W. "Deciding which queue to join: some counterexamples", *Oper. Res.*, **1**, pp 55-62 (1986).
8. Menich, R. and Serfozo, R.F. "Optimality of routing and servicing in dependent parallel processing systems", *Queueing Systems*, **9**, pp 403-418 (1991).
9. Hordijk, A. and Koole, G. "On the optimality of the generalized shortest queue policy", *Prob. Eng. Info. Sci.*, **4**, pp 477-487 (1990).
10. Towsley, D., Sparaggis, P.D. and Cassandras, C.G. "Optimal routing and buffer allocation for a class of finite capacity queueing networks", *Proc. 29th Conf. on Decision and Control*, pp 658-663 (1990).
11. Sparaggis, P.D., Towsley, D. and Cassandras, C.G. "Extremal properties of the shortest/longest non-full queue policies in finite capacity systems with state-dependent service

rates", *J. of Appl. Prob.*, **30**, pp 223-236 (1993).

12. Ancker, C.J. and Gafarian, A.V. "Some queueing problems with balking and reneging", *Oper. Res.*, **11**, pp 88-100 (1963).
13. Ancker, C.J. and Gafarian, A.V. "Queueing with impatient customers who leave at random", *J. of Industrial Eng.*, **3**, pp 86-87 (1962).
14. Baccelli, F., Boyer, P. and Hebuterne, G. "Single-server queues with impatient customers", *Adv. Appl. Prob.*, **16**, pp 887-905 (1984).
15. Barrer, D.Y. "Queueing with impatient customers and ordered service", *Oper. Res.*, **5**, pp 650-656 (1957).
16. Daley, D.J. "General customer impatience in queue GI/G/1", *J. of Appl. Prob.*, **2**, pp 186-205 (1965).
17. Takacs, L. "A single-server queue with limited virtual waiting time", *J. of Appl. Prob.*, **11**, pp 612-617 (1974).
18. Bhattacharya, P.B. and Ephremides, A. "Optimal scheduling with strict deadlines", *IEEE Trans. Automat. Contr.*, **7**, pp 721-728 (1989).
19. Panwar, S.S., Towsley, D. and Wolf, J.K. "Optimal scheduling policies for a class of queues with customer deadlines to the beginning of service", *J. of ACM*, **4**, pp 832-844 (1988).
20. Bellman, R.A. "A Markovian decision process", *J. of Math. Mech.*, **6**, pp 679-684 (1957).
21. Howard, R. *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, USA (1960).
22. Marshall, A.W. and Olkin, I. *Inequalities: Theory of Majorization and Its Applications*, Academic Press, Inc., New York, USA (1979).
23. Lippman, S. "Applying a new device in the optimization of exponential queueing systems", *Oper. Res.*, **23**, pp 687-710 (1957).