OPTIMAL ROUTING IN SYSTEMS WITH ILR SERVICE TIME DISTRIBUTIONS

Don Towsley and Panayotis D. Sparaggis

CMPSCI Technical Report 93-13
February 1993

Optimal Routing in Systems with ILR Service Time Distributions[1]

Don Towsley, Department of Computer Science
Panayotis D. Sparaggis, Department of Electrical and Computer Engineering

University of Massachusetts
Amherst, MA 01003

ABSTRACT

We consider the problem of routing jobs to one of K parallel queues. Arrivals are independent of the state of the system but otherwise arbitrary. Assuming that queues have infinite capacities and the service times form a sequence of i.i.d. random variables with Increasing Likelihood Rate (ILR) distribution, we prove that the Shortest Queue (SQ) policy minimizes the vectors of queue lengths in the sense of weak Schur-convex ordering. We give a counterexample which shows that this result is not generally true when the service times have Increasing Hazard Rate (IHR) but are not increasing in the likelihood rate sense. Finally, we show that when capacities are finite, the SQ policy stochastically maximizes the departure process and minimizes the loss counting process.

February 1993

Submitted to the *Journal of Applied Probability*

[1]This work was partially supported by NSF under contract NCR-9116183 and by an IBM Graduate Fellowship Award.

1 Introduction

A classical problem in the literature of control of queues is the determination of the optimal routing policy for customers that arrive in front of a set of K parallel *homogeneous* service stations with infinite or finite buffer capacities. The assumption of homogeneity refers to the fact that all of the customers' service times are independent and identically distributed (i.i.d.) random variables. Arrivals are independent of the state of the system, but otherwise arbitrary, and the service discipline is FIFO. When the service times are exponential, it has been shown that the *Shortest Queue* (SQ) policy minimizes queue length vectors in the sense of *weak Schur-convex ordering.* (e.g., Winston [13] and Ephremides et. al. [1]). When the capacities at the stations are finite, the optimality of the SQ policy extends to the minimization of the number of losses that occur by any time t, again provided that the service times form i.i.d. sequences of exponential random variables (see Hordijk and Koole [2], Menich and Serfozo [5], and Towsley et. al. [9]). When service times are not exponential, Whitt [12] shows that it is not always optimal to join the shortest queue.

In this paper, we show that the SQ policy is, in fact, optimal for a much broader class of service time distributions; namely, distributions with *Increasing Likelihood Ratio* (ILR) (e.g. [3]). Our results cover both infinite and finite capacity systems. Specifically, we prove that SQ minimizes the number of customers in the system in the sense of a weak Schur-convex ordering. Moreover, SQ stochastically maximizes the departure counting process and also, when there are finite buffers, it minimizes the loss counting process. We give a simple counterexample (the main idea drawn from Righter and Shanthikumar [6]) which shows that when service times have *Increasing Hazard Rate* (IHR), but are not increasing in the likelihood rate sense, SQ need not be optimal. This contradicts a result by Weber [10] stating that SQ stochastically maximizes the departure process when service times have IHR.

Our arguments involve the construction of auxiliary policies that allow for idling and/or deliberate rejection of customers, and that are compared against an arbitrary policy π on a sample path. The main idea is to show that, given π, one can construct a sequence of policies (starting from π), such that each policy reverses the routing decision when the previous one in the sequence violates the SQ rule for the first time, resulting in a monotonically decreasing sequence of queue length vectors in the sense of majorization. The construction is described in section 3 which treats systems with infinite capacities, following some preliminary results on stochastic orderings that are given in section 2. Section 4 contains the counterexample for IHR distributions. Finally, the extension to finite-buffer systems is given in section 5.

2 Preliminaries

Let $\mathbf{N}, \mathbf{M}$, be two K-dimensional real-valued vectors. We introduce the notation $\hat{N}_k$ to denote the *k-th largest element* in vector $\mathbf{N}$ and define the following ordering (see [4]).

Definition 1 *Vector* $\mathbf{N}$ *is said to majorize vector* $\mathbf{M}$ *(written* $\mathbf{M} \prec \mathbf{N}$*) if*

$$\sum_{i=1}^{k} \hat{N}_i \geq \sum_{i=1}^{k} \hat{M}_i, \quad k = 1, \cdots, K-1,$$

$$\sum_{i=1}^{K} \hat{N}_i = \sum_{i=1}^{K} \hat{M}_i \ . \tag{1}$$

The definition of majorization describes the fact that the elements of $\mathbf{M}$ are 'less spread-out' than the elements of $\mathbf{N}$; equivalently, $\mathbf{M}$ is 'more balanced' than $\mathbf{N}$. A weaker ordering can be defined by replacing the equality in (1) by an inequality. This implies, $\sum_{i=1}^{k} \hat{N}_i \geq \sum_{i=1}^{k} \hat{M}_i$, $k = 1, \cdots, K$. In this case, vector $\mathbf{N}$ is said to *weakly submajorize* vector $\mathbf{M}$ (or, weakly majorize $\mathbf{M}$ from below) written $\mathbf{M} \prec_w \mathbf{N}$. Functions related to majorization are defined as follows.

Definition 2 *A function* $\phi : I\!N^K \rightarrow I\!R$ *is said to be Schur-convex iff*

$$\mathbf{M} \prec \mathbf{N} \ \Rightarrow \ \phi(\mathbf{M}) \leq \phi(\mathbf{N}), \ \forall \mathbf{M}, \mathbf{N} \in I\!N^K.$$

If ϕ is increasing and Schur-convex then $\mathbf{M} \prec_w \mathbf{N} \Rightarrow \phi(\mathbf{M}) \leq \phi(\mathbf{N})$. Marshall and Olkin [4] define the following stochastic ordering between random vectors.

Definition 3 *If* $\mathbf{N}$ *and* $\mathbf{M}$ *are random vectors, we have*

$$\mathbf{M} \leq_{E_1^\uparrow} \mathbf{N} \ \textit{if} \ E[\phi(\mathbf{M})] \leq E[\phi(\mathbf{N})], \ \forall \phi \ \textit{increasing and Schur-convex.}$$

This is also called the weak Schur-convex ordering. Recall that the definition of common stochastic ordering among two random variables is, $X \leq_{st} Y$ if $Ef(X) \leq Ef(Y)$ for all increasing f. Furthermore, a process $\{X(t); t \geq 0\}$ is said to be stochastically smaller than another process $\{Y(t); t \geq 0\}$ if

$$(X(t_1), ..., X(t_n)) \leq_{st} (Y(t_1), ..., Y(t_n)),$$

for all $n, t_1, ..., t_n$. Similarly, $\{\mathbf{N}(t); t \geq 0\} \leq_{E_1^\uparrow} \{\mathbf{M}(t); t \geq 0\}$ if

$$(\mathbf{N}(t_1), ..., \mathbf{N}(t_n)) \leq_{E_1^\uparrow} (\mathbf{M}(t_1), ..., \mathbf{M}(t_n))$$

for all $n, t_1, ..., t_n$. For counting processes, a stronger sample path stochastic ordering can be defined as follows (e.g., [11]).

Definition 4 *If* $\{X(t); t \geq 0\}$, $\{Y(t); t \geq 0\}$ *are counting processes then* $\{X(t); t \geq 0\} \supseteq \{Y(t); t \geq 0\}$ *iff there exist on a common probability space two counting processes* $\{\tilde{X}(t); t \geq 0\}$ *and* $\{\tilde{Y}(t); t \geq 0\}$ *that are equal in law to* $\{X(t); t \geq 0\}$ *and* $\{Y(t); t \geq 0\}$ *respectively, such that for all time intervals* $[a, b]$ *the jump epochs of* $\{\tilde{Y}(t); t \geq 0\}$ *are a subset of those of* $\{\tilde{X}(t); t \geq 0\}$.

Clearly, $\{X(t); t \geq 0\} \supseteq \{Y(t); t \geq 0\} \Rightarrow \{X(t); t \geq 0\} \geq_{st} \{Y(t); t \geq 0\}$. Finally, we recall the definitions of non-negative random variables that have increasing hazard rate (IHR) and increasing likelihood rate (ILR) distributions. Let X be a random variable with density f and distribution function F. We say that X has increasing hazard rate if $f(t)/(1-F(t))$ is increasing in t. On the other hand, we say that X is increasing in likelihood ratio if $f(t_1+a)/f(t_2+a)$ is increasing in a for all $t_1 \leq t_2$. It is well known that ILR distributions include truncated normal, uniform, exponential, Poisson and geometric distributions. Furthermore, it is true that if X is ILR, then it is IHR.

3 Optimality of the SQ policy

In this section, we establish the optimality of the SQ policy when the service times have ILR distributions and the service stations have infinite capacities. Let Σ denote the class of policies that have instantaneous information regarding the queue lengths and the *elapsed* service times of the customers in service. Let $N_i^\pi(t)$ denote the number of customers at queue i at time t, given a policy π in Σ, and $\mathbf{N}^\pi(t) = (N_1^\pi(t), ..., N_K^\pi(t))$. For the customer that occupies the server at the ith station, let $x_i^\pi(t)$ be its elapsed service time and $s_i^\pi(t)$ be its remaining service time. Finally, let $D^\pi(t)$ denote the number of departures by time t under π.

Throughout the paper, we say that queue i is longer than queue j at time t if this is true in the sense of Weber [10], i.e., $N_i^\pi(t) > N_j^\pi(t)$, or, $N_i^\pi(t) = N_j^\pi(t)$ and $x_1^\pi(t) \leq x_2^\pi(t)$. Equivalently, we say that j is shorter than i. Let SQ be the policy that always routes to the shortest queue. Let X_t denote the remaining lifetime of a random variable X from time t on, given that it exceeds t. In order to prove the optimality of the SQ policy, we need the following result that is shown as part of a construction in [3].

Lemma 1 *Let $X, Y \in I\!R^+$ be two continuous random variables that have the same density function f and are increasing in likelihood ratio. Then for all $s < t$, (Y_s, Y_t) is equal in law to $(\tilde{Y}_s, \tilde{Y}_t)$ that is defined as follows.*

$$\begin{aligned}(\tilde{Y}_s, \tilde{Y}_t) &= 1(X_s \leq X_t)(X_t, X_s) + \\ &\quad +1(X_s > X_t)[U(X_s, X_t)(X_t, X_s) + (1 - U(X_s, X_t))(X_s, X_t)], \qquad (2)\end{aligned}$$

where $U(X_s, X_t)$ is a Bernoulli r.v. that is equal to 0 with probability $p(X_s, X_t)$, equal to 1 otherwise, $p(X_s, X_t) = [f(t+X_s)f(s+X_t)]/[f(t+X_t)f(s+X_s)]$.

Note that for $s < t$ and $X_s > X_t$ it follows that $p(X_s, X_t) \leq 1$ due to the assumption of an ILR distribution. We now have the following result.

Theorem 1 *In a symmetric routing system with ILR service stations SQ will minimize the vector of queue lengths in the sense of $E_1^\uparrow$ ordering, i.e.,*

$$\{\mathbf{N}^{SQ}(t); t \geq 0\} \leq_{E_1^\uparrow} \{\mathbf{N}^\pi(t); t \geq 0\}, \ \forall \pi \in \Sigma,$$

provided that $\mathbf{N}^{SQ}(0) =_{st} \mathbf{N}^\pi(0)$.

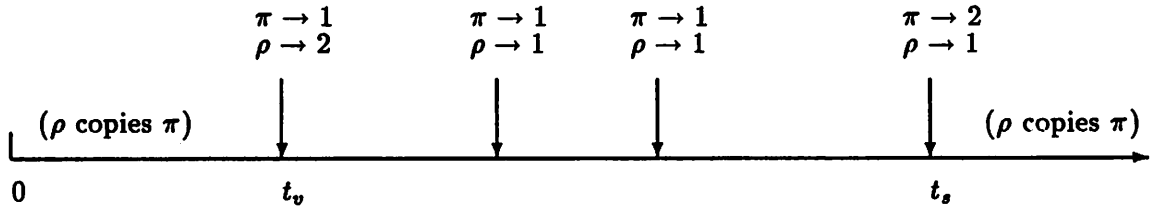

Figure 1: Arrivals and routing decisions under π and ρ; t_s is a synchronization point.

Proof. It suffices to prove the theorem for $K = 2$. We condition on arrival times and initial queue lengths. We will construct a policy ρ, that allows for *idling* such that the following is true on any sample path:

$$\mathbf{N}^{\rho}(t) \prec \mathbf{N}^{\pi}(t), \ \forall t \geq 0. \tag{3}$$

We say that the two systems (i.e., the one under ρ and the one under π) are *synchronized* at any time t if $(\hat{N}_1^{\rho}(t), \hat{N}_2^{\rho}(t)) = (\hat{N}_1^{\pi}(t), \hat{N}_2^{\pi}(t))$, and the customer in service at the queue with the larger (smaller) queue length (in the sense of Weber) under ρ has the same elapsed or the same remaining time as the one at the queue with the larger (resp. smaller) queue length under π. The proof is by forward induction on event times, i.e., arrival and departure times, starting at $t = 0$ at which (3) holds trivially by assumption.

Initially, ρ copies π until the first time, say t_v, that π violates the SQ rule, at which time ρ reverses the routing decision. Suppose without loss of generality that at time t_v^- queue 1 is larger than queue 2 (in the sense of Weber), in which case π routes to queue 1, whereas ρ routes to queue 2. Between time t_v and the time the two systems synchronize, we define ρ to route to queue 1 whenever π routes to queue 1 and to idle queue 2 whenever $N_2^{\pi}(t) = 0$. The first time π routes to queue 2 (before we get synchronization) we define ρ to route to queue 1, in which case we will prove that the two systems synchronize[2]. After synchronization occurs, ρ simply copies π[3]. An example is shown in Figure 1.

By a straightforward coupling of the service times, we get $\mathbf{N}^{\rho}(t) = \mathbf{N}^{\pi}(t)$, $0 \leq t < t_v$. Our proof starts by observing that the following relations are true at time t_v.

$$(x_1^{\pi}(t_v), x_2^{\pi}(t_v)) = (x_1^{\rho}(t_v), x_2^{\rho}(t_v)), \tag{4}$$

$$N_1^{\pi}(t_v) = N_1^{\rho}(t_v) + 1; \ \ N_2^{\pi}(t_v) = N_2^{\rho}(t_v) - 1; \ \ N_1^{\pi}(t_v) > N_2^{\pi}(t_v), \tag{5}$$

$$\mathbf{1}[N_1^{\pi}(t_v) = N_2^{\pi}(t_v) + 1] = 1 \ \Rightarrow \ x_1^{\pi}(t_v) < x_2^{\pi}(t_v). \tag{6}$$

Note in particular how (6) holds by assumption: $\mathbf{1}[N_1^{\pi}(t_v) = N_2^{\pi}(t_v) + 1] = 1$ implies $N_1^{\pi}(t_v^-) = N_2^{\pi}(t_v^-)$ (since π routes to queue 1 at t_v); thus , since π violates the SQ rule, $x_1^{\pi}(t_v^-) < x_2^{\pi}(t_v^-)$.

[2]In this case our construction essentially reduces to an interchange argument.

[3]Possibly under relabeling of the two queues, since the definition of synchronization allows for a permutation of the queues.

Our objective is to prove that equation (5) (which implies the majorization $\mathbf{N}^{\rho}(t_v) \prec \mathbf{N}^{\pi}(t_v)$) does, in fact, propagate along event times until the first synchronization point after time t_v.

Clearly, equations (4)-(6) will remain true if an arrival occurs at queue 1 (under both ρ and π by our construction). If an arrival occurs at queue 2 under π, then ρ will route to queue 1, in which case it is straightforward to see that the two systems will synchronize. Finally, if a service completion occurs, say at time t_s, equations (4)-(6) will again remain true, provided that $N_1^{\pi}(t_s^-) > N_2^{\pi}(t_s^-) + 1$. This last restriction will ensure that, if the service completion is at queue 1, then $N_1^{\pi}(t_s) > N_2^{\pi}(t_s)$ (which is needed to ensure (5) at t_s). Note that if $N_1^{\pi}(t_s^-) = N_2^{\pi}(t_s^-) + 2$ then (6) will hold true at t_s if the service completion is at queue 1, since in that case $x_1^{\pi}(t_s) = 0 < x_2^{\pi}(t_s)$.

Let us now consider the only remaining case, i.e., the case in which (4)-(6) hold at time some t_s^-, with $N_1^{\pi}(t_s^-) = N_2^{\pi}(t_s^-) + 1$ and t_s is a service completion event time. The situation at t_s^- is as follows.

$$(x_1^{\pi}(t_s^-), x_2^{\pi}(t_s^-)) = (x_1^{\rho}(t_s^-), x_2^{\rho}(t_s^-)), \tag{7}$$

$$N_1^{\pi}(t_s^-) = N_2^{\rho}(t_s^-); \;\; N_2^{\pi}(t_s^-) = N_1^{\rho}(t_s^-); \;\; N_1^{\pi}(t_s^-) = N_2^{\pi}(t_s^-) + 1. \tag{8}$$

$$x_1^{\pi}(t_s^-) < x_2^{\pi}(t_s^-). \tag{9}$$

In this case, we cannot couple the queues in the two systems directly as above. For example, if a service completion occurred at queue 1 under both π and ρ, then the majorization ordering would be violated since then, $N_1^{\pi}(t_s) = N_2^{\pi}(t_s)$, $N_2^{\rho}(t_s) = N_1^{\pi}(t_s) + 1$, $N_1^{\rho}(t_s) = N_2^{\pi}(t_s) - 1$. However, based on Lemma 1, we can cross-couple the remaining service times of the customers at the two queues, so that at t_s one of the following is true:

T1 The service completion is at queue 1 under ρ and queue 2 under π and $s_2^{\rho}(t_s) = s_1^{\pi}(t_s)$.

T2 The service completion is at queue 2 under both ρ and π and $s_1^{\rho}(t_s) = s_1^{\pi}(t_s)$.

T3 The service completion is at queue 2 under ρ and queue 1 under π and $s_1^{\rho}(t_s) = s_2^{\pi}(t_s)$.

Let p_2 be the probability that T2 occurs, given that one of T2 or T3 occurs, and $p_3 = 1 - p_2$. Using the notation of Lemma 1, $p_2 = \int_0^{\infty}[1 - p(x_1^{\rho}(t_s^-) + s, x_2^{\rho}(t_s^-))]g_{s|x_1^{\rho}(t_s^-)}(s)ds$, where $g_{s|a}(s)$ is the conditional density function of the residual service time, given that the service time exceeds a. After the service completion, i.e., at time t_s^+, policy ρ acts as follows.

A1 If the service completion is at queue 1, then ρ assumes that the service completion occurred at queue 2 under π.

A2 If the service completion is at queue 2, then, with probability p_2, ρ assumes that the service completion is at queue 2 under π.

A3 If the service completion is at queue 2, then, with probability p_3, ρ assumes that the service completion is at queue 1 under π.

We couple A1-A3 with T1-T3 above. If T1 occurs, the two systems will synchronize since $N_i^\pi(t_s) = N_{(i+1)mod2}^\rho(t_s)$ and $s_2^\rho(t_s) = s_1^\pi(t_s)$, $x_1^\rho(t_s) = x_2^\pi(t_s) = 0$. The case for T3 is similar. On the other hand, if T2 occurs, the situation at t_s is as follows.

$$(s_1^\pi(t_s), x_2^\pi(t_s)) = (s_1^\rho(t_s), x_2^\rho(t_s)), \tag{10}$$

$$N_1^\pi(t_s) = N_1^\rho(t_s) + 1; \;\; N_2^\pi(t_s) = N_2^\rho(t_s) - 1; \;\; N_1^\pi(t_s) > N_2^\pi(t_s) + 1. \tag{11}$$

Note that (10), (11) are similar to (4) and (5) respectively (with (11) being, in fact, slightly stronger than (5)). Clearly, the above two equations remain true if an arrival occurs at queue 1 (under both π and ρ), or if a departure occurs at queue 2. If an arrival occurs at queue 2 under π, then ρ will route to queue 1 and the two systems will synchronize. Finally, if a departure occurs at queue 1, the two systems will go back to a state described by equations (4)-(6).

Thus, since (5) (sometimes in the stronger form of (11)) does in fact hold until synchronization occurs, (3) has been shown. Using a straightforward coupling of service times, it is easy to see that, given ρ, there exists a policy ρ' that routes exactly where ρ does, but does not allow for idling, so that $N_i^{\rho'}(t) \leq N_i^\rho(t)$, $i = 1, 2$, for all $t \geq 0$. Thus,

$$\mathbf{N}^{\rho'}(t) \prec_w \mathbf{N}^\rho(t) \prec \mathbf{N}^\pi(t), \;\; t \geq 0.$$

Repeating this construction, it then follows that SQ minimizes queue lengths in the sense of weak majorization, i.e., $\mathbf{N}^{SQ}(t) \prec_w \mathbf{N}^\pi(t)$, $\forall t \geq 0$. Since

$$\mathbf{N}_1 \prec_w \mathbf{M}_1 \text{ on } I\!R^K, \;\; \mathbf{N}_2 \prec_w \mathbf{M}_2 \text{ on } I\!R^K \;\Rightarrow\; (\mathbf{N}_1, \mathbf{N}_2) \prec_w (\mathbf{M}_1, \mathbf{M}_2) \text{ on } I\!R^{2K},$$

it follows that $(\mathbf{N}^{SQ}(t_1), ..., \mathbf{N}^{SQ}(t_n)) \prec_w (\mathbf{N}^\pi(t_1), ..., \mathbf{N}^\pi(t_n))$ for all $n, t_1, ..., t_n$, which implies the desired result. ∎

Remark. During the construction of ρ, we required that the density function of the service times is known. This is needed, however, only after the first time π violates the SQ rule. Since our arguments involve the construction of a sequence of policies that eventually reach the SQ policy (that, of course, never violates the SQ rule), knowing the density function is ultimately not necessary: SQ simply routes to the shortest queue, without taking into account the form of the density function.

Note that in the sample path considered in the proof, it is true that $D^\rho(t) = D^\pi(t)$, for all $t \geq 0$, and, of course, $D^\rho(t) \leq D^{\rho'}(t)$ since under ρ' customers do not delay due to idling. Thus, we have the following result.

Corollary 1 *In a symmetric routing system with ILR service stations SQ will stochastically maximize the departure process, i.e.,*

$$\{D^{SQ}(t); t \geq 0\} \geq_{st} \{D^\pi(t); t \geq 0\} \;\; \forall \pi \in \Sigma,$$

provided that $\mathbf{N}^{SQ}(0) = \mathbf{N}^\pi(0)$.

4 Counterexample for IHR stations

In this section we show that if the stations have IHR service time distributions, then it is not necessarily true that $D^{SQ}(t) \geq_{st} D^{\pi}(t)$ for all $t \geq 0$. Our counterexample uses a distribution that was first used by Righter and Shanthikumar [6] to show a limitation of IHR distributions in proving the throughput optimality of FIFO over the class of preemptive policies in open queueing networks.

We consider a system consisting of two parallel stations, each having one customer in the queue initially. We assume that time is discrete and that the service distribution is geometric with parameter 1/2, truncated at 3. That is, a customer requires 1, 2, or 3 units of times of service with probability 1/2, 1/4 and 1/4, respectively. This is an IHR, but not an ILR, distribution. We further assume that, at time zero, the customer at queue 1, say C_1, has already received one unit of service, and the customer at queue 2, say C_2 has received no service. Suppose now that at time 0 a new customer, say C_3, arrives. SQ routes C_3 to queue 1, whereas another policy π routes C_3 to queue 2. We assume that no more arrivals occur in the system.

It is easy to see that the probability of having exactly three departures by the end of the second time unit is greater under π than SQ. More specifically, the probability of $D^{SQ}(2)$ being equal to 3 is equal to the product of the probabilities that each of C_1 and C_3 finish in one time unit (each equal to 1/2), and the probability that C_2 requires no more than 2 time units (equal to 3/4). On the other hand, the probability of $D^{\pi}(2)$ being equal to 3 is only 1/4, equal to the product of the probabilities that each of C_2 and C_3 finish in exactly one time slot (by assumption C_1 requires no more than two time units). It is well known that $X \leq_{st} Y$ if and only if $Pr[X > a] \leq Pr[Y > a]$ for all a (e.g. [7]). This is an equivalent definition to the one given in section 2. Since no more than three departures are possible by the end of the second time unit, $Pr[D^{SQ}(2) = 3] < Pr[D^{\pi}(2) = 3]$ implies $D^{SQ}(t) \not\geq_{st} D^{\pi}(t)$.

The intuition behind the counterexample is simple. Since both C_1 and C_2 are equally likely to finish in exactly one time slot, it is useful to route C_3 to queue 2, since, as C_1 needs no more than two time units of service, one effectively requires that only C_2 and C_2 finish in exactly one time slot in order to have three customers departed by the end of the second time slot.

5 Extension to finite buffers

In this section we consider systems in which customers have ILR service time distributions and service stations have *finite* capacities. Let Σ_f be the class the policies that have information regarding the queue lengths and the elapsed service times of the customers in service, and are required to route a customer to a queue that has available space, if one exists. Let $L^{\pi}(t)$ denote the number of customers that are rejected and lost due to insufficient buffer space by time t under a policy π in Σ_f. We have the following result.

Theorem 2 *In a symmetric routing system with ILR finite-capacity service stations, SQ will stochastically maximize the departure process and minimize the loss process, i.e.,*

$$\{(D^{SQ}(t), -L^{SQ}(t)); t \geq 0\} \geq_{st} \{(D^{\pi}(t), -L^{\pi}(t)); t \geq 0\} \ \forall \pi \in \Sigma_f,$$

provided that $\mathbf{N}^{SQ}(0) =_{st} \mathbf{N}^{\pi}(0)$.

Proof. Following the same construction as in Theorem 1, it is seen that $\{(D^{\rho}(t), -L^{\rho}(t)); t \geq 0\} =_{st} \{(D^{\pi}(t), -L^{\pi}(t)); t \geq 0\}$. Thus, following the arguments in the proof of Theorem 1, it remains to show that there exists a policy ρ' such that

$$\{(D^{\rho'}(t), -L^{\rho'}(t)); t \geq 0\} \geq_{st} \{(D^{\rho}(t), -L^{\rho}(t)); t \geq 0\}. \tag{12}$$

Let us first define ρ' as in Theorem 1, where now, in addition, we require that ρ' deliberately rejects a customer, even if there exists available space, when ρ does. Thus, $\rho \notin \Sigma_f$. Clearly, $\mathbf{N}^{\rho'}(t) \prec_w \mathbf{N}^{\rho}(t)$ for all $t \geq 0$, from which follows that ρ' results in the same number of losses as ρ and a larger number of departures than ρ on the sample path. Therefore, we have shown (12). To complete the proof, we must relax the assumption that ρ' may deliberately reject a customer.

Let ρ'' be the policy that routes exactly where ρ' does but without allowing for deliberate rejections. Thus, $\rho'' \in \Sigma_f$. Let $\{A_i^{\gamma}; t \geq 0\}$ denote the arrival counting process at queue i, under $\gamma = \rho', \rho''$. Clearly, $\{A_i^{\rho''}; t \geq 0\} \supseteq \{A_i^{\rho'}; t \geq 0\}$, $i = 1, 2$. Using Lemma 1 from [8] it follows that $\{(D^{\rho''}(t), -L^{\rho''}(t)); t \geq 0\} \geq_{st} \{(D^{\rho'}(t), -L^{\rho'}(t)); t \geq 0\}$, which finally yields the desired result. ∎

References

[1] A.Ephremides, P.Varaiya and J.Walrand, 'A simple dynamic routing problem', *IEEE Trans. on Aut. Control*, vol. AC-25, 1980.

[2] A.Hordijk and G.Koole, 'On the optimality of the generalized shortest queue policy', *Probability in the Eng. and Info. Sciences*, vol. 4, pp. 477-487, 1990.

[3] Z.Liu and D.Towsley, 'Stochastic scheduling in in-forest networks', COINS Technical Report, U. of Massachusetts, March 1992.

[4] A.W.Marshall and I.Olkin, *Inequalities: Theory of majorization and its applications*, Academic Press, 1979.

[5] R.Menich and R.F.Serfozo, 'Optimality of routing and servicing in dependent parallel processing systems', *Queueing Systems*, vol. 9, pp. 403-418, 1991.

[6] R.Righter and J.G.Shanthikumar, 'Extremal properties of the FIFO discipline in queueing networks', preprint, 1991.

[7] S.M.Ross, *Stochastic processes*, Wiley, New York, 1983.

[8] P.D.Sparaggis and C.G.Cassandras, 'Throughput monotonicity in communication networks with blocking: properties and counterexamples', *1991 IEEE INFOCOM*, pp. 2C.4.1-2C.4.10, 1991.

[9] D.Towsley, P.D.Sparaggis and C.G.Cassandras, 'Optimal routing and buffer allocation for a class of finite capacity queueing systems', *IEEE Trans. on Aut. Control*, vol. 37, pp. 1446-1451, 1992.

[10] R.R.Weber, 'On the optimal assignment of customers to parallel queue', *J. of Applied Prob.*, vol. 15, pp. 406-413, 1978.

[11] W.Whitt, 'Comparing counting processes and queues', *Advances in Applied Prob.*, vol. 13, pp. 207-220, 1981.

[12] W.Whitt, 'Deciding which queue to join', *Oper. Research*, vol. 34, pp. 55-62, 1986.

[13] W.Winston, 'Optimality of the shortest line discipline', *J. of Applied Prob.*, vol. 14, pp. 181-189, 1977.