Minimizing response times and queue lengths in systems of parallel queues*

Ger Koole
Department of Mathematics and Computer Science, Vrije Universiteit,
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Panayotis D. Sparaggis
Department of Electrical and Computer Engineering, University of Massachusetts,
Amherst, MA 01003, U.S.A.

Don Towsley
Department of Computer Science, University of Massachusetts,
Amherst, MA 01003, U.S.A.

Published in *Journal of Applied Probability* **36**:1185-1193, 1999

Abstract

We consider the problem of routing customers to one of two 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 Ratio (ILR) distribution, we prove that the Shortest Queue (SQ) policy minimizes various cost functionals related to queue lengths and response times. We give a counterexample which shows that this result is not generally true when the service times have Increasing Hazard Rate 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.
Keywords: Routing; stochastic majorization; response times; increasing likelihood ratio distributions
AMS subject classification: 60K25

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 capacity queues. 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 disciplines at the queues are FIFO. When the service times are exponentially distributed it has been shown that the *Shortest Queue (SQ)* policy minimizes queue length vectors in the sense of *weak Schur*

*This work was partially supported by NSF under contract NCR-9116183, by an IBM Graduate Fellowship Award, and by a European Human Capital and Mobility grant.

convex ordering (e.g., Winston [13] and Ephremides et al. [2]). 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 exponentially distributed random variables (see Hordijk and Koole [3], Menich and Serfozo [6], and Towsley et al. [10]). When service times are not exponentially distributed 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. [4]). Moreover, we consider various performance measures, namely queue lengths, response times, and the number of departures. Our results cover both infinite and finite capacity systems, but we have to restrict ourselves to two queues. Specifically, we prove that the SQ policy minimizes the number of customers in the system in the sense of a weak Schur convex ordering. The SQ policy also minimizes the vector of response times in a separable weak Schur convex ordering. Moreover, the SQ policy stochastically maximizes the departure counting process and, when there are finite buffers, minimizes the loss counting process. We give a simple counterexample (the main idea drawn from Righter and Shanthikumar [8]) that 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 [11] stating that the SQ policy stochastically maximizes the departure process when service times have IHR.

Our arguments involve the construction of auxiliary policies that allow for the idling and/or the deliberate rejection of customers, and their comparison 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 model and some preliminary results on stochastic orderings are given in Section 2. The construction (for infinite capacity systems) is described in Section 3. The results for the various performance measures can be found in Section 4. (The reader who is only interested in the results can skip Section 3.) Section 5 contains the counterexample for IHR distributions. Finally, extensions to finite capacity systems are given in Section 6.

2 The model

We are concerned with the assignment of n customers, which have arrival times $0 < t_1 < \ldots < t_n$. Let Σ denote the class of routing policies that have, at time t, instantaneous information regarding the queue lengths and the *elapsed* service times of the customers in service from 0 to t. Let $N_k^\pi(t)$ denote the number of customers at queue k at time t, given a policy $\pi \in \Sigma$, and $N^\pi(t) = (N_1^\pi(t), N_2^\pi(t))$. Let $x_k^\pi(t)$ be the elapsed service time of the customer that occupies the server at the kth station (we take $x_k^\pi(t) = 0$ if $N_k^\pi(t) = 0$). The (two-dimensional) vector $(N^\pi(t), x^\pi(t))$ can be seen as the state of the system at t. The policies in Σ are also allowed to *idle* the servers. This means that, as long as server k idles, $x_k^\pi(t)$ does not increase.

Throughout the paper, we say that queue k is longer than queue j at time t if this is true in the sense of Weber [11], i.e., $N_k^\pi(t) > N_j^\pi(t)$, or, $N_k^\pi(t) = N_j^\pi(t)$ and $x_k^\pi(t) \leq x_j^\pi(t)$. Equivalently, we say that j is shorter than k. Let SQ be the non-idling policy that always routes to the shortest queue.

Let S denote the service time of a customer and S_t denote the remaining service time of S given that it has received t time units of service, i.e., $P(S_t \leq s) = P(S \leq s + t \mid S > t)$. Assume for the moment that S_t has a density f_t. A service time is said to be ILR (Increasing in Likelihood Ratio)

if, for $r_1 < r_2$ and $t_1 \geq t_2$, $f_{t_1}(r_1)f_{t_2}(r_2) \geq f_{t_1}(r_2)f_{t_2}(r_1)$ (see e.g. [4]). For discrete distributions we should replace $f_t(r)$ by $P(S_t = r)$. Examples of ILR service times include those that are constant and exponentially distributed. The following result holds ([7]):

Lemma 2.1 *If S is ILR and $t_1 \geq t_2$ then*

$$P(S_{t_1} = r_1 \mid \min\{S_{t_1}, S_{t_2}\} = r_1, \max\{S_{t_1}, S_{t_2}\} = r_2) \geq \frac{1}{2}$$

for all relevant $r_1 < r_2$.

3 The construction

In this section we construct a sequence of policies starting from an arbitrary policy π which converges to SQ. This is done using coupling arguments. This construction is used in the next section. If a performance measure increases (decreases) for each pair of policies in the above list, then it is maximized (minimized) by SQ.

Consider some non-idling policy π^i which always routes customers $i, \ldots, n$ according to SQ, but customer $i-1$ differently, for some or all arrival time and service time sequences. We construct a policy γ (which may idle a server) that routes the first $i-2$ customers to the same queues as π^i and routes customers $i-1, \ldots, n$ to the shortest queue. We then construct another policy γ' by removing the inserted idleness in γ. This is the main step in the construction: the repeated use of this step allows us to obtain a sequence of policies starting from an arbitrary policy π that converges to SQ.

We now go into the details of this construction.

We consider a particular realization of the system under γ. (Thus it would be notationally correct to add the realization ω to all variables; we refrain from doing in order to maintain a simple notation.) At the same time, using the same service time realizations as γ, we construct a copy of the system under π^i, which is in its turn used by γ to make its decisions. We will see, while making the construction, that this can be done and that it results in a stochastically correct copy of the system under π^i. The properties of the construction are used in the next section to prove our results.

We identify the states that the systems can occupy under γ and under the constructed copy of π^i. Up to time t_{i-1} we couple the systems such that they behave the same. This is possible because the routing decisions (and the initial states) are identical for $t < t_{i-1}$. Thus for $t < t_{i-1}$ the following holds:

Case 1 (synchronization). $N_k^{\pi^i} = N_k^{\gamma}$ and $x_k^{\pi^i} = x_k^{\gamma}$, for $k = 1, 2$.

Now consider the routing decision at t_{i-1}. If, for this particular realization, π^i routes the arriving customer to the shortest queue, then we remain in case 1. If not, then without loss of generality we assume that π^i routes customer $i-1$ to queue 2, while γ routes it to queue 1.

Given that SQ routes the customer to queue 1 and π^i routes it to queue 2, either $N_1^{\pi^i}(t_{i-1}^-) < N_2^{\pi^i}(t_{i-1}^-)$ or $N_1^{\pi^i}(t_{i-1}^-) = N_2^{\pi^i}(t_{i-1}^-)$ and $x_1^{\pi^i}(t_{i-1}^-) > x_2^{\pi^i}(t_{i-1}^-)$, and therefore at t_{i-1} we are in one of the two following situations.

Case 2. $N_2^{\pi^i} = N_2^{\gamma} + 1$, $N_1^{\gamma} = N_1^{\pi^i} + 1$, $N_1^{\pi^i} < N_2^{\gamma}$, and $x_k^{\pi^i} = x_k^{\gamma}$ where $x_1^{\bullet}$ and $x_2^{\bullet}$ are arbitrary.

Case 3. $N_2^{\pi^i} = N_2^{\gamma} + 1$, $N_1^{\gamma} = N_1^{\pi^i} + 1$, $N_1^{\pi^i} = N_2^{\gamma} > 0$, and $x_k^{\pi^i} = x_k^{\gamma}$ with $x_1^{\bullet} > x_2^{\bullet}$.

The construction method described below can give rise to two additional situations after t_{i-1}.

Case 4. $N_2^{\pi^i} = N_2^{\gamma} + 1$, $N_1^{\gamma} = N_1^{\pi^i} + 1$, $N_1^{\pi^i} < N_2^{\gamma}$, $x_1^{\pi^i} = x_1^{\gamma}$, $x_1^{\bullet} < x_2^{\pi^i}, x_2^{\gamma}$, and the departure epoch of the customers present in queue 2 are already coupled, i.e., they depart at the same time.

Case 5. $N_k^{\pi^i} = N_k^\gamma$, $x_1^{\pi^i} = x_1^\gamma$, $x_1^\bullet < x_2^{\pi^i}, x_2^\gamma$, and the customers present in queue 2 depart at the same time due to the coupling.

Next, we consider each of these 5 cases and show that each possible event yields a transition from one case to another. This is done by observing the sample path of γ, waiting for events to occur. If the event is a departure from either queue 1 or 2, as described below, we couple a departure (at possibly a different queue) under π^i such that the marginal distribution of the systems remain correct. In all cases, both π^i and γ route according to SQ.

Case 1. We couple the arrival and departure events such that the systems remain synchronized.

Case 2. First consider an arrival under γ. Recall that both π^i and γ apply the SQ rule. Thus under π^i the arrival is sent to queue 1. For γ this depends on the queue lengths and the elapsed service times. It can be seen that transitions to case 1, 2, and 3 are possible.

If the first event to occur under γ is a departure, then we couple to this a departure in the same queue under π^i. As $x^{\pi^i} = x^\gamma$, the marginal distributions remain correct. The only exception we make is for queue 1: if $N_1^\gamma = 1$, then we force the server to idle because $N_1^{\pi^i} = 0$. After the departure the systems enter states that are covered by either case 2 or case 3.

Case 3. Consider an arrival under γ. Then π^i routes it to queue 1. Also under γ the customer is routed to queue 2 and the systems become synchronized. Note that the customer cannot be routed to queue 2, as π^i follows SQ.

The departures are coupled in a somewhat complex way, and this constitutes the crucial step in the proof where the assumption that the service time distribution is ILR is used. The departures from both queues must be considered together. Assume that under the current sample path ω the first event to happen is a departure, with realization $\min\{S_{x_1}, S_{x_2}\}(\omega) = 0$. We immediately sample $\max\{S_{x_1}, S_{x_2}\}$, which has value $r > 0$. Now we consider two different sample paths at the same time, those where under γ the departure is from queue 1 and the remaining service time for the customer at queue 2 is r, and vice versa. Using the property described in lemma 2.1, and taking $p = P(S_{x_1} = 0 \mid \min = 0, \max = r)$, we find that $p \geq 1/2$ (where the probability measure is conditional on the fact that $\min\{S_{x_1}, S_{x_2}\}$ is the first event to happen). Thus we can couple the departures as follows:

i) with probability $1 - p$ departures at queue 2 under π^i and at queue 1 under γ occur;
ii) with probability $1 - p$ departures at queue 1 under π^i and at queue 2 under γ occur;
iii) with probability $2p - 1$ a customer leaves queue 1 in both systems.

Observe that the remaining service times associated with the customers in the remaining two queues are coupled as well, and the customers depart after r time units. For the first possibility we arrive at systems with $N_1^{\pi^i} = N_2^{\pi^i} = N_1^\gamma = N_2^\gamma$ with $x_2^{\pi^i} = x_1^\gamma = 0$ and $x_1^{\pi^i} \neq x_2^\gamma$, but due to the coupling these customers leave at the same time. Thus case 5 applies, with queues 1 and 2 interchanged under π^i. The second possiblity leads also to case 5 again with queues 1 and 2 interchanged under π^i. The third possibility leads to case 4.

Case 4. An arrival is always sent to queue 1 under π^i. Under γ, it is sent to queue 2 if $N_1^{\pi^i} + 1 = N_2^\gamma$ (leading to case 5), and to queue 1 otherwise (in which case we remain in case 4). A departure at queue 1 leads again to case 4. A departure at queue 2 leads to either case 2 or 3 as the coupled customers disappear.

Case 5. Arrivals and departures at the same queues are coupled; case 1 applies as soon as the first customers in queue 2 are gone. Note that although $x_2^\gamma \neq x_2^{\pi^i}$, the assignments (according to SQ) are the same because $x_1^\bullet < x_2^{\pi^i}, x_2^\gamma$.

So far the construction has led us from the policy π^i to γ, which allows a server to idle when there is work in its associated queue. Using a straightforward coupling argument of service times it is easy to see that there exists a policy γ' that routes exactly as γ does but does not allow servers

to idle, so that either case 1 applies, or the following:

Case 6. For $k = 1, 2$, either $N_k^{\gamma'} = N_k^{\gamma}$ and $x_k^{\gamma'} \geq x_k^{\gamma}$ or $N_k^{\gamma'} < N_k^{\gamma}$ with $x_k^{\gamma'}, x_k^{\gamma}$ arbitrary.

Lemma 3.1 *Any cost functional which is decreasing in some ordering when passing from π^i to γ to γ' is minimized by SQ.*

In the next section we give several examples of such functionals.

Remark. Crucial to the construction is the fact that both π^i and γ use SQ from customer i on. Indeed, the coupling for case 3 depends on the maximum of two service times and thus uses future information. Therefore the policies are not allowed to depend on the coupling. This is where the construction breaks down if we have more than 2 queues: in that case we need γ to deviate from SQ to remain in one of our 5 cases.

4 Results

4.1 Queue lengths

We start with results related to the queue lengths. To do so we need to introduce the majorization ordering. Let N, M be two K-dimensional real-valued vectors. We introduce the notation $N_{[k]}$ to denote the kth largest element of N and define the following ordering (see [5]).

Definition 4.1 *A vector N is said to majorize vector M (written $M \prec N$) if*

$$\sum_{j=1}^{k} N_{[j]} \geq \sum_{j=1}^{k} M_{[j]}, \ k = 1, \cdots, K,$$

with equality for $k = K$.

The definition of majorization describes the fact that the elements of M are 'less spread-out' than the elements of N; equivalently, M is 'more balanced' than N. If we omit the condition on the equality for $k = K$ we define the *weak majorization* ordering $\prec_w$. If $M \prec_w N$ than M is smaller and/or more balanced than N.

A check of all 6 cases of the previous section along with the realization that the (weak) majorization ordering is independent of the ordering of the components gives the following.

Theorem 4.2 *In a routing system with two homogeneous ILR service stations the queue lengths under* SQ *are weakly majorized by the queue length of any other policy, i.e.,*

$$\{N^{\mathrm{SQ}}(t); t \geq 0\} \prec_w \{N^{\pi}(t); t \geq 0\}, \ \forall \pi \in \Sigma,$$

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

Note that the construction of the previous section ensures that $N^{\gamma} \prec N^{\pi^i}$ and $N^{\gamma'} \leq N^{\gamma}$, yielding $N^{\gamma'} \prec_w N^{\pi^i}$.

The result holds jointly over all t due to the fact that we coupled the sample paths during the construction.

Every function of the queue lengths which is increasing in the weak majorization ordering is minimized by SQ. These functions are called the *weak Schur convex functions*.

Definition 4.3 *A function $\phi : \mathbb{N}^K \to \mathbb{R}$ is said to be (weak) Schur convex iff*

$$M \prec (\prec_w) N \Rightarrow \phi(M) \leq \phi(N),\ \forall M, N \in \mathbb{N}^K.$$

It can be shown that a function is weak Schur convex iff it is increasing in all components and Schur convex. Examples of weak Schur convex functions are $\phi(N) = \sum_k N_k$ and $\phi(N) = \max_k N_k$. Now let us introduce the *stochastic ordering* $\leq_s$. Two r.v.s X and Y have $X \leq_s Y$ if $P(X > w) \leq P(Y > w)$ for all w. The following result can be derived from Proposition 8.2.2 of [9]. We use $\stackrel{\mathrm{d}}{=}$ to denote "equally distributed".

Lemma 4.4 *For rvs X and Y, $X \leq_s Y$ holds iff X' and Y' can be constructed such that $X \stackrel{\mathrm{d}}{=} X'$ and $Y \stackrel{\mathrm{d}}{=} Y'$ and X' and Y' can be coupled (i.e., they live on the same probability space) with $X'(\omega) \leq Y'(\omega)$ for each $\omega \in \Omega$.*

It follows from the construction described in the previous section that $\phi(N^{\mathrm{SQ}}(t)) \leq \phi(N^{\pi}(t))$ for each realization whenever ϕ is weak Schur convex, and therefore $\phi(N^{\mathrm{SQ}}(t)) \leq_s \phi(N^{\pi}(t))$. Thus SQ *stochastically* minimizes any weak Schur convex cost function ofthe queue lengths. Furthermore, the proof is by a *pathwise* construction (jointly over all t), and thus SQ stochastically minimizes expressions of the form $\left(\phi_1(N^{\pi}(s_1)), \ldots, \phi_M(N^{\pi}(s_M))\right)$ (with all ϕ_m weak Schur convex) as well (see Section 8.2 of [9]).

4.2 Departure process

We now consider a set of cost functions associated with the departure process. Let $D^{\pi}(t)$ denote the number of departures by time t under π. Note that the following is true of the sample paths constructed in Section 3, $D^{\gamma}(t) = D^{\pi^i}(t)$, for all $t \geq 0$, and of course, $D^{\gamma}(t) \leq D^{\gamma'}(t)$ since under γ'_i customers do not delay due to idling. This yields the following result.

Corollary 4.5 *In a symmetric routing system with two ILR service stations, SQ stochastically maximizes the departure process, i.e.,*

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

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

Note that in all our results we can replace $N^{\mathrm{SQ}}(0) = N^{\pi}(0)$ by $N^{\mathrm{SQ}}(0) \stackrel{\mathrm{d}}{=} N^{\pi}(0)$, as we can couple the systems at $t = 0$.

4.3 Response times

Let us now consider the response times of the customers. It is clear that the response time of an arbitrary customer i, R_i, is not minimized by sending the customers before it to the shortest queue; therefore we have to consider the *vector* of response times. Let us consider the construction in detail. It is clear that $R^{\gamma}_j \stackrel{\mathrm{d}}{=} R^{\pi^i}_j$, for $j \leq i-2$ and $j > i$, and that $R^{\gamma}_{i-1} \leq_s R^{\pi^i}_{i-1}$. From the construction it follows that $R^{\gamma}_{i-1} + R^{\gamma}_i \stackrel{\mathrm{d}}{=} R^{\pi^i}_{i-1} + R^{\pi^i}_i$. Therefore $(R^{\gamma}_{i-1}, R^{\gamma}_i) \prec (R^{\pi^i}_{i-1}, R^{\pi^i}_i)$ in distribution.

The second step of the construction gives $R^{\gamma'}_k \leq R^{\gamma}_k$ for all k. Concluding, we find that $R^{\gamma'}_k \leq_s R^{\pi^i}_k$ for all $k \neq i-1, i$, and that $(R^{\gamma'}_{i-1}, R^{\gamma'}_i) \prec_w (R^{\pi^i}_{i-1}, R^{\pi^i}_i)$ in distribution.

This statement is of the form $R_k^\gamma \leq_s R_k^{\pi^i}$, and not of the form $R_j^\gamma \leq R_j^{\pi^i}$. This is due to the fact that we coupled the queue lengths, but not the response times of the individual customers. Thus we have no results on the joint distribution of the response times; therefore the allowable cost functional need to be separable. It is easily verified that the class of separable weak Schur convex functions is equal to the class of functions of the form $\phi(N) = \sum_k g(N_k)$, with g increasing and convex.

Theorem 4.6 *In a routing system with two homogeneous ILR service stations the response times under SQ are minimized in the separable weak Schur convex ordering, provided that* $N^{\mathrm{SQ}}(0) = N^\pi(0)$.

Remark. Theorem 4.6 can easily be shown to hold for an arbitrary number of homogeneous service stations provided that service times are restricted to be exponentially distributed rvs.

5 Counterexample for IHR stations

The distribution function of a non-negative rv S (with density f and distribution function F) is IHR (increasing in hazard rate) if $f(t)/(1-F(t))$ is increasing in t. This assumes that S has a density; a more general definition is the following (with $\bar{F}(t) = 1 - F(t)$): S is IHR if $\bar{F}(t+s)/\bar{F}(t)$ is decreasing in $-\infty < t < \infty$ with $\bar{F}(t) > 0$, for each $s \geq 0$ ([1], p. 54).

In this section we show that if the stations have IHR service time distributions then it is not necessarily true that $D^{\mathrm{SQ}}(t) \geq_s D^\pi(t)$ for all $t \geq 0$. Our counterexample uses a distribution that was first used by Righter and Shanthikumar [8] 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 additional 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 $D^{\mathrm{SQ}}(2)$ 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). Since no more than three departures are possible by the end of the second time unit, $Pr(D^{\mathrm{SQ}}(2) = 3) < Pr(D^\pi(2) = 3)$ implies $D^{\mathrm{SQ}}(t) \not\geq_s D^\pi(t)$. This contradicts a result in [11] stating that SQ stochastically maximizes the departure process when service times have IHR.

The intuition behind the counterexample is simple. Since both C_1 and C_3 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_3 finish in exactly one time slot in order to have three customers departed by the end of the second time slot.

6 Extension to finite buffers

In this section we consider systems in which customers have ILR service time distributions and service stations have *equal finite* capacities. Customers routed to a full queue are lost. We allow the policies in Σ to reject arriving customers, even if there is space in one or both queues available. SQ still routes to the shortest queue and does not reject customers, unless both queues are full.

We adapt the construction of Section 3 to the finite capacity case. We assume that π^i does not unnecessarily reject customers. It is readily seen that γ blocks customers when π^i does. When comparing γ and γ' it can occur that under γ customers are blocked which can be admitted under γ'; we reject these customers. Now we construct another policy γ''. Let i be the last customer which is (for some realizations) voluntarily rejected by γ'. The policy γ'' is equal to γ' up to t_i, and assigns customer i to some non-full queue j. After that it uses its copy of γ' to behave the same. When under γ' an arrival is assigned to queue j then synchronization occurs. If before that moment the extra customer in queue j is ready to be served (i.e., all other customers have left), then the server idles. This gives us as seventh case. Let $L^\pi(t)$ denote the number of customers that are rejected or lost by time t under policy $\pi \in \Sigma$.

Case 7. $N_k^{\gamma''} = N_k^{\gamma'}$ for $k \neq j$, $N_j^{\gamma''} = N_j^{\gamma'} + 1$, $x_k^{\gamma''} = x_k^{\gamma}$ for $k = 1, 2$, and $L^{\gamma'} = L^{\gamma''} + 1$.

Obviously, using the results of Section 3, we can remove the inserted idleness, and then use the above construction again to avoid other rejections, etc.

Let us consider one by one the cost functionals discussed in Section 4.

6.1 Queue lengths

From case 7 it is clear that Theorem 4.2 does not hold anymore. However, it is easily seen that cost functionals as $\sum_k N_k + L$ are minimized by SQ.

6.2 Departure process

Now let $D^\pi(t)$ denote the number of departures from one of the queues by time t under π. The following is readily seen.

Theorem 6.1 *In a routing system with two homogeneous ILR service stations that have finite equal capacities SQ stochastically maximizes the departure process and minimizes the loss process, i.e.,*

$$\{(D^{\mathrm{SQ}}(t), -L^{\mathrm{SQ}}(t)); t \geq 0\} \geq_s \{(D^\pi(t), -L^\pi(t)); t \geq 0\} \ \forall \pi \in \Sigma,$$

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

6.3 Response times

If we take $R_i = \infty$ if customer i is rejected then Theorem 4.6 still holds. To come to a more useful result however, we could take for such a customer $R_i = R$ with R some large number. Unfortunately the step from γ' to γ'' need no longer lead to an improvement, as the first customer joining queue j after customer i under γ' will have a stochastically lower response time than customer i under γ''.

Acknowledgment

The authors are thankful to Rhonda Righter for pointing out an error in the original construction for the model with more than two queues.

References

[1] R.E. Barlow and F. Proschan. *Statistical Theory of Reliability and Life Testing*. Holt, Rinehart and Winston, 1975.

[2] A. Ephremides, P. Varaiya, and J. Walrand. A simple dynamic routing problem. *IEEE Transactions on Automatic Control*, 25:690–693, 1980.

[3] A. Hordijk and G.M. Koole. On the optimality of the generalized shortest queue policy. *Probability in the Engineering and Informational Sciences*, 4:477–487, 1990.

[4] Z. Liu and D. Towsley. Stochastic scheduling in in-forest networks. *Advances in Applied Probability*, 26:222–241, 1994.

[5] A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, 1979.

[6] R. Menich and R.F. Serfozo. Monotonicity and optimality of symmetric parallel processing systems. *Queueing Systems*, 9:403–418, 1991.

[7] R. Righter. Scheduling. In M. Shaked and J.G. Shanthikumar, editors, *Stochastic Orders and their Applications*, pages 381–432. Academic Press, 1994.

[8] R. Righter and J.G. Shanthikumar. Extremal properties of the FIFO discipline in queueing networks. *Journal of Applied Probability*, 29:967–978, 1992.

[9] S.M. Ross. *Stochastic Processes*. Wiley, 1983.

[10] D. Towsley, P.D. Sparaggis, and C.G. Cassandras. Optimal routing and buffer allocation for a class of finite capacity queueing systems. *IEEE Transactions on Automatic Control*, 37:1446–1451, 1992.

[11] R.R. Weber. On the optimal assignment of customers to parallel queues. *Journal of Applied Probability*, 15:406–413, 1978.

[12] W. Whitt. Deciding which queue to join: Some counterexamples. *Operations Research*, 34:55–62, 1986.

[13] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, 14:181–189, 1977.