Optimality of the Round Robin Routing Policy[*]

Zhen LIU[†]
INRIA Centre Sophia Antipolis
2004 Route des Lucioles
06560 Valbonne
France

Don TOWSLEY
Dept. of Computer Science
University of Massachusetts
Amherst, MA 01003
U.S.A.

August 10, 1992, Revised February 11, 1993

Abstract

In this paper we consider the problem of routing customers to identical servers, each with its own infinite capacity queue. Under the assumptions that *i)* the service times form a sequence of independent and identically distributed random variables with increasing failure rate distribution and *ii)* state information is not available, we establish that the round robin policy minimizes, in the sense of a separable increasing convex ordering, the customer response times and the numbers of customers in the queues.

Keywords : Optimal Routing, Round Robin, Sample Path Analysis, Stochastic Ordering.

Submitted to *J. Appl. Prob.*

[*] This work was supported in part by the National Science Foundation under grant ASC 88-8802764 and NCR-9116183.

[†] The work of this author was also partially supported by CEC DG-XIII under the ESPRIT-BRA grant QMIPS.

1 Introduction

Consider a stream of customers arriving to a controller which immediately routes them to one of several identical infinite capacity single server queues. We establish the optimality of the round robin (RR) routing policy when the service times are independent and identically distributed (i.i.d.) random variables (r.v.'s) having an increasing failure rate (IFR) distribution, provided the controller has available to it the past routing decisions but no queue length information. More precisely, we prove that the RR policy minimizes the customer response times and the numbers of customers in the queues in the sense of separable increasing convex ordering (see definition below).

When the controller has available to it the vector of workloads of the queues, it has been established by various authors (see Kingman [12], Vasicek [23], Foss [6, 7], Wolff [28, 29] and Daley [4]) that the Smallest Workload (SW) policy, which routes the arriving customer to the queue with lowest workload, minimizes the stationary waiting times in the sense of increasing convex ordering. Transient optimality results of the SW policy were obtained by Foss [6, 7], Wolff [28, 29] and Daley [4] for the (e.g. Kiefer-Wolfowitz) workload vector, and by Liu and Towsley [13, 14] for customer response times.

When the controller has available to it the vector of queue lengths, it has been shown by Winston [27], Weber [25] and Menich [16] (see also Walrand [24]) that the shortest queue (SQ) policy minimizes the number of customers in system when the service times are exponentially distributed. When the service times have an increasing likelihood ratio distribution, Towsley and Sparaggis [22] have shown that the SQ policy is still optimal provided that the server whose customer has been in service the longest is chosen in the case of a tie. Whitt [26] provided counterexamples showing that SQ is not optimal in general. Several studies [10, 21, 19] have established the optimality of the SQ policy for finite buffer queues, under the assumption that service times are exponential random variables. Menich and Serfozo [17] have considered the joint routing/scheduling problem when there is an additional (movable) server. They established the optimality of the SQ/LQ (LQ stands for the policy serving longest queue). In the case of finite buffers, the duality between various routing and scheduling problems where queue lengths are available to the controller has been established in [18].

When the controller has available to it only the past routing decisions but no queue length information, it has been proved by Ephremides et al. [5] (see also Hajek [9] and Walrand [24]) that the RR policy is optimal when service times are exponentially distributed. When the service times have general distribution, Stoyan [20] and Jean-Marie and Liu [11] have shown that the RR policy yields smaller (in the sense of increasing convex ordering) stationary and transient, respectively, customer waiting times than the Bernoulli policy with equal routing probability for each queue. This last Bernoulli routing policy has been shown to be optimal among all the Bernoulli routing policies [2, 3, 8]. Our results can be considered to be extensions of the results in [5, 11, 20].

The paper is organized as follows. In the next section, we provide a detailed description of the model and state the main result. The proof of this result is found in Section 3. Comments on the result are given in Section 4.

2 Formal Model and Main Result

2.1 Model Description

There are $s > 1$ parallel queues, each with its own infinite capacity waiting buffer and a single server. The servers are identical and have the same speed, say 1. Customers arrive at a controller which immediately routes them to one of the s queues. The service discipline of each queue is FCFS. The system is initially empty. (Note that our results actually hold in the case where the initial queue lengths are i.i.d. r.v.'s).

The controller has available to it the past routing decisions and all arrival times, but no queue length information. We denote the class of such routing policies as Σ.

Denote by a_n and σ_n the arrival time and the service time of customer n, respectively, $n \geq 1$, with $a_1 = 0 < a_2 < \cdots < a_n < \cdots$. The sequence of service times $\{\sigma_n\}_{n=1}^{\infty}$ consists of i.i.d. r.v.'s having an IFR distribution. The sequence of arrival times $\{a_n\}_{n=1}^{\infty}$ is independent of the service times, but is otherwise arbitrary. In particular, it can be a deterministic sequence.

We are interested in one member of Σ, namely the round robin policy which will be referred to as ρ. Let π_n denote the identity of the server that customer n is routed to by policy $\pi \in \Sigma$. Then the round robin policy is defined to be the policy that selects servers cyclically beginning with server 1, i.e., $\rho_n = n - s\lfloor (n-1)/s \rfloor$, $n = 1, 2, \cdots$, where $\lfloor x \rfloor$ denotes the integer part of x.

Let $\pi \in \Sigma$ be an arbitrary routing policy. Let c_n^π be the completion time of customer n under π, and d_n^π be the n-th departure time in the s queues under π. Denote by $N_i^\pi(t)$ the number of customers at server $i = 1, \cdots, s$ including the one in service at time $t > 0$ under policy π and $\boldsymbol{N}^\pi(t) = (N_1^\pi(t), \cdots, N_s^\pi(t))$. Similarly, denote by $U_i^\pi(t)$ the unfinished work at server

$i = 1, \cdots, s$ at time $t > 0$ under policy π and $\boldsymbol{U}^{\pi}(t) = (U_1^{\pi}(t), \cdots, U_s^{\pi}(t))$. Last, denote by R_n^{π} the response time of the n-th customer in the system under π and $\boldsymbol{R}^{\pi}(m) = (R_1^{\pi}, \cdots, R_m^{\pi})$, $m = 1, 2, \cdots$. Here

$$R_n^{\pi} = c_n^{\pi} - a_n, \quad n = 1, 2, \cdots.$$

2.2 Stochastic Ordering

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors in $I\!R^m$. Following the notation of [15], we define $\mathcal{C}_3^{\uparrow}$ to be the class of functions: $f: I\!R^m \to I\!R$ of the form $f(\boldsymbol{X}) = \sum_{i=1}^m g(X_i)$, where $g: I\!R \to I\!R$ is increasing and convex. We say that $\boldsymbol{X} \leq_{E_3^{\uparrow}} \boldsymbol{Y}$ if

$$E[f(X)] \leq E[f(Y)], \qquad \forall f \in \mathcal{C}_3^{\uparrow},$$

provided the expectations exist. The ordering $\leq_{E_3^{\uparrow}}$ is referred to as separable increasing convex ordering in this paper. When dimension $m = 1$, $\leq_{E_3^{\uparrow}}$ coincides with the increasing convex ordering $\leq_{icx}$ between random variables.

The proof of our results uses the notion of majorization. Let $\boldsymbol{x}, \boldsymbol{y} \in I\!R^m$ be two real vectors. Vector $\boldsymbol{x}$ is said to be weakly majorized by vector $\boldsymbol{y}$ (written $\boldsymbol{x} \prec_w \boldsymbol{y}$) iff

$$\sum_{j=1}^{k} x_{[j]} \leq \sum_{j=1}^{k} y_{[j]}, \quad k = 1, \cdots, m,$$

where the notation $x_{[j]}$ is taken to be the j-th largest element of x. Various properties concerning these notions can be found in [15].

2.3 Optimality of the Round Robin Routing Policy

The main result of the paper can be stated as follows.

Theorem 1 *Assume the system is empty at time $t = 0$. Assume that service times form a sequence of i.i.d. r.v.'s with IFR distribution. Then for all $\pi \in \Sigma$*

$$\boldsymbol{N}^{\rho}(t) \;\leq_{E_3^{\uparrow}}\; \boldsymbol{N}^{\pi}(t), \quad t \geq 0, \tag{1}$$

$$\boldsymbol{R}^{\rho}(m) \;\leq_{E_3^{\uparrow}}\; \boldsymbol{R}^{\pi}(m), \quad m = 1, 2, \cdots. \tag{2}$$

3 Proof of the Optimality of the Round Robin Routing Policy

In order to prove Theorem 1, we construct the queue lengths under policies ρ and π on the same probability space. Under this construction, the joint queue length statistics will not be preserved under either policy. However, the marginal behavior of every queue will be statistically correct. It turns out that such a construction collapses to the one used in [5] when service times are exponentially distributed r.v.'s.

Let $\widehat{\Omega}$ be such probability space and $\mathcal{S}(\pi, \rho)$ be the system composed of s queues controlled by policy π and s queues controlled by policy ρ. We assume that the initial queue lengths are zero under both policies. The arrival times are $a_1, a_2, \cdots$. At a_n, $n = 1, 2, \cdots$, a customer is routed to one of the first s queues according to policy π, and another customer is routed to one of the other s queues according to policy ρ.

Let $\{u_k\}_{k=1}^{\infty}$ be a sequence of i.i.d. r.v.'s uniformly distributed in the interval $[0, 1)$. We shall refer to these as the service parameters associated with $\mathcal{S}$.

Let $F(x) = \Pr[\sigma \leq x]$ denote the cumulative distribution of the service time. Let σ_t denote the remaining time of σ given that it exceeds t and

$$F_t(x) \overset{\text{def}}{=} \Pr[\sigma_t \leq x] = \frac{F(t+x) - F(t)}{1 - F(t)}, \qquad x \geq 0; \quad t \geq 0. \tag{3}$$

An event in $\mathcal{S}(\pi, \rho)$ corresponds to an arrival or a service completion under either π or ρ. Let e_k, $k = 1, 2, \cdots$, be the time epoch of the k-th event in $\mathcal{S}(\pi, \rho)$. The system $\mathcal{S}(\pi, \rho)$ is constructed in such a way that at event time e_k, $k = 1, 2, \cdots$, all customers in service under both policies have their service times recomputed according to u_k and the remaining service time distributions. A customer having been already in service for v units of time will receive a new remaining service time $F_v^{-1}(u_n)$.

To be more precise, we define the following notation and provide a set of recursive equations. Let $\phi \in \{\pi, \rho\}$ be an arbitrary policy.

- $\widehat{N}_i^{\phi}(t)$ denotes the queue length (including the customer in service) of queue i at time $t \geq 0$ under ϕ. The process $\widehat{N}_i^{\phi}(t)$ will be assumed to be right-continuous in t.

- $r_i^{\phi}(t)$ denotes the remaining service time of the customer in service, if any, at queue i at time t under ϕ. If $\widehat{N}_i^{\phi}(t) = 0$, then $r_i^{\phi}(t) = 0$.

- $w_i^\phi(t)$ denotes the time at which the customer in service (if any) commenced its service. If $\widehat{N}_i^\phi(t^-) = 0$, then $w_i^\phi(t) = t$.

- $\hat{s}_n^\phi$, $\hat{d}_n^\phi$ denote, respectively, the times of the of the n-th service commencement and the n-th departure in the s queues controlled by ϕ.

- $\hat{I}_s^\phi(n)$, $\hat{I}_d^\phi(n)$ denote the identities of the n-th customer that starts service and finishes service, respectively, in the s queues controlled by ϕ.

- χ_k identifies the k-th event. If $\chi_k = 0$, then the k-th event is an arrival. Otherwise, $\chi_k = 1$ indicates that the k-th event is a service completion under either or both policies.

- A_k denotes the number of arrivals in the interval $[0, e_k]$.

These variables are defined by the recursive equations as follows, where $\mathbf{1}(\cdot)$ is the indicator function.

$$\begin{aligned}
e_{k+1} &= \min\left(a_{A_k+1},\ \left\{\min_{\phi\in\{\pi,\rho\}}\ \min_{\{1\le i\le s,\ \widehat{N}_i^\phi(e_k)>0\}}\ r_i^\phi(e_k)\right\}\right), \\
\widehat{N}_i^\phi(e_{k+1}) &= \widehat{N}_i^\phi(e_k) + \mathbf{1}\left(\chi_k = 0 \wedge \phi_{A_k+1} = i\right) - \mathbf{1}\left(\chi_k = 1 \wedge r_i^\phi(e_k) = e_{k+1} - e_k\right), \\
w_i^\phi(e_{k+1}) &= \mathbf{1}\left(\widehat{N}_i^\phi(e_k) > 0 \wedge e_{k+1} - e_k < r_i^\phi(e_k)\right) w_i^\phi(e_k) \\
&\quad + \mathbf{1}\left(e_{k+1} - e_k = r_i^\phi(e_k) \vee \widehat{N}_i^\phi(e_k) = 0\right) e_{k+1}, \\
r_i^\phi(e_{k+1}) &= F^{-1}_{e_{k+1} - w_i(e_{k+1})}(u_{k+1})\mathbf{1}\left(\widehat{N}_i^\phi(e_{k+1}) > 0\right), \\
\chi_{k+1} &= \mathbf{1}\left(\left\{\min_{\phi\in\{\pi,\rho\}}\ \min_{\{1\le i\le s,\ \widehat{N}_i^\phi(e_{k+1})>0\}}\ r_i^\phi(e_{k+1})\right\} < a_{A_k+1} - e_k\right), \\
A_{k+1} &= A_k + \mathbf{1}(\chi_{k+1} = 0).
\end{aligned}$$

Note that the construction couples the remaining service times of all of the customers in service in such a way that the remaining service times are determined by the distribution functions of their remaining service times and a common uniformly distributed random variable. Since the service times have the same IFR distribution, the resulting virtual service completion times are in the same order as the service commencement times. To be more precise,

Lemma 1 *For any $\phi, \phi' \in \{\pi, \rho\}$, and for any $m, n = 1, 2, \cdots$,*

(i) if $\hat{b}^{\phi}_{m} \leq \hat{b}^{\phi'}_{n}$, then $\hat{c}^{\phi}_{m} \leq \hat{c}^{\phi'}_{n}$, where $\hat{b}^{\phi}_{n}$ is the service commencement time of customer n under policy ϕ;

(ii) $\hat{I}^{\phi}_{d}(n) = \hat{I}^{\phi}_{s}(n)$.

Proof. As the distribution function F is IFR, we know that

$$F_u(x) \leq F_v(x), \qquad 0 \leq u < v, \quad x \geq 0.$$

Therefore, from the monotonicity of the functions F_u and F_v, we get

$$F_u^{-1}(x) \geq F_v^{-1}(x), \qquad 0 \leq u < v, \quad 0 \leq x \leq 1.$$

Using the above relation, it is easy to prove property *(i)* by induction on the event epochs e_k, $k = 1, 2, \cdots$. The detailed proof is omitted for the sake of brevity.

Property *(ii)* is an immediate consequence of property *(i)*. ∎

The construction of system $\mathcal{S}(\pi, \rho)$ also exhibits the following important properties.

Lemma 2 *The following properties hold for $\mathcal{S}(\pi, \rho)$,*

(a) for all $n \geq 1$, $\hat{b}^{\rho}_{n} = \hat{s}^{\rho}_{n}$, $\hat{c}^{\rho}_{n} = \hat{d}^{\rho}_{n}$;

(b) for all $n \geq 1$, $\hat{s}^{\rho}_{n} \leq \hat{s}^{\pi}_{n}$, $\hat{d}^{\rho}_{n} \leq \hat{d}^{\pi}_{n}$;

(c) $|\widehat{N}^{\rho}_{i}(t) - \widehat{N}^{\rho}_{j}(t)| \leq 1, \quad i, j = 1, \cdots, s, \quad t \geq 0.$

Proof. The key to establishing the first two properties is the following relation:

$$\hat{s}^{\phi}_{n} \geq \max(a_n, \hat{d}^{\phi}_{n-s}), \qquad \phi \in \{\pi, \rho\}, \quad n = 1, 2, \cdots. \tag{4}$$

(Here it is understood that $\hat{d}^{\phi}_{n} = 0$ for $n \leq 0$.) This relation is established by induction. The basis step is easy as $\hat{s}^{\phi}_{1} = a_1 \geq \hat{d}^{\phi}_{1-s} = 0$. Consider the inductive step. Assume that for some $m \geq 1$, the relation holds for all $n \leq m$. We establish it for $n = m + 1$. There are two cases depending on whether or not $a_{m+1} \geq \hat{d}^{\phi}_{m-s+1}$. If $a_{m+1} \geq \hat{d}^{\phi}_{m-s+1}$ then clearly $\hat{s}^{\phi}_{m+1} \geq a_{m+1} \geq \hat{d}^{\phi}_{m-s+1}$, which establishes the relation. Assume instead that $\hat{d}^{\phi}_{m-s+1} > a_{m+1}$ and that the relation is false, so that $\hat{s}^{\phi}_{m+1} < \hat{d}^{\phi}_{m-s+1}$. According to Lemma 1, the s customers $\hat{I}^{\phi}_{d}(m-s+1), \cdots, \hat{I}^{\phi}_{d}(m)$ are served by at most $s-1$ *distinct* servers. This implies the existence of some customer, say j, that started at time $\hat{s}^{\phi}_{l}$ where $\hat{s}^{\phi}_{l} \geq \hat{d}^{\phi}_{m-s+1} > \hat{s}^{\phi}_{m+1}$ and completed at time $\hat{d}^{\phi}_{l}$ where $\hat{d}^{\phi}_{m+1} > \hat{d}^{\phi}_{l} > \hat{s}^{\phi}_{l}$. This contradicts property *(i)* of Lemma 1. Therefore the relation holds and the inductive step is complete.

Observe now that under the round robin policy,

$$\hat{b}^{\rho}_{n} = \max(a_n, \hat{c}^{\rho}_{n-s}), \quad n = 1, 2, \cdots. \tag{5}$$

We prove by induction that for all $n = 1, 2, \cdots$,

$$\hat{b}^{\rho}_{n} = \hat{s}^{\rho}_{n}, \quad \hat{c}^{\rho}_{n} = \hat{d}^{\rho}_{n} \tag{6}$$

and that

$$\hat{s}^{\rho}_{n} = \max(a_n, \hat{d}^{\rho}_{n-s}). \tag{7}$$

Clearly, for $n = 1$, the above equations hold. Assume they are true for some $m \geq 1$. Consider $n = m + 1$. By the inductive assumption,

$$\hat{s}^{\rho}_{m+1} \leq \hat{b}^{\rho}_{m+1} = \max(a_{m+1}, \hat{c}^{\rho}_{m+1-s}) = \max(a_{m+1}, \hat{d}^{\rho}_{m+1-s}).$$

It then follows from (4) that relation (7) holds for $m + 1$ and that $\hat{s}^{\rho}_{m+1} = \hat{b}^{\rho}_{m+1}$. Using further property *(i)* of Lemma 1 implies $\hat{d}^{\rho}_{m+1} = \hat{c}^{\rho}_{m+1}$. Hence, (6) holds for $m + 1$. Therefore, by induction, relations (6) and (7) hold for all $n \geq 1$, so that property *(a)* is valid.

We prove now property *(b)* by induction. The basis step in the induction proof is easy to establish as $\hat{s}^{\rho}_{1} = \hat{s}^{\pi}_{1} = a_1$. Furthermore, according to Lemma 1, the first customer served will be the first customer completed under both policies. According to the construction, they receive the same service time under both policies; hence $\hat{d}^{\rho}_{1} = \hat{d}^{\pi}_{1}$.

Assume that property *(b)* holds up to $n = m$. For $n = m + 1$, we have,

$$\begin{aligned} \hat{s}^{\rho}_{m+1} &= \max(a_{m+1}, \hat{d}^{\rho}_{m-s+1}), && \text{relation (7)} \\ &\leq \max(a_{m+1}, \hat{d}^{\pi}_{m-s+1}), && \text{by induction,} \\ &\leq \hat{s}^{\pi}_{m+1}, && \text{relation (4),} \end{aligned}$$

so that, by Lemma 1, $\hat{d}^{\rho}_{m+1} \leq \hat{d}^{\pi}_{m+1}$. By induction, property *(b)* holds for all $n \geq 1$.

Consider the third property. Let $K_i(t)$ and $L_i(t)$, $i = 1, \cdots, s$, denote, respectively, the number of arrivals and completions at queue i under ρ by time t in the system $\mathcal{S}(\pi, \rho)$. Consider queues i and j where $i < j$. We have, according to the definition of the round robin policy,

$$K_j(t) + 1 \geq K_i(t) \geq K_j(t).$$

Furthermore, due to property *(a)*,

$$L_j(t) + 1 \geq L_i(t) \geq L_j(t).$$

Therefore,

$$\begin{aligned} N^{\rho}_i(t) - N^{\rho}_j(t) &= (K_i(t) - K_j(t)) - (L_i(t) - L_j(t)) \\ &\leq 1 - (L_i(t) - L_j(t)) \\ &\leq 1, \end{aligned}$$

and

$$\begin{aligned} N^{\rho}_i(t) - N^{\rho}_j(t) &= (K_i(t) - K_j(t)) - (L_i(t) - L_j(t)) \\ &\geq -(L_i(t) - L_j(t)) \\ &\geq -1, \end{aligned}$$

so that property *(c)* holds.

This completes the proof of the lemma. ■

Finally, we prove the claim we made previously about the preservation of the marginal distribution of the queue lengths in $\mathcal{S}(\pi, \rho)$. In what follows, $=_d$ denotes equality in distribution.

Lemma 3 *For any $\phi \in \{\pi, \rho\}$, the following equalities are true,*

$$\widehat{N}^{\phi}_i(t) \;=_d\; N^{\phi}_i(t), \qquad i = 1, \cdots, s; \quad t \geq 0. \tag{8}$$

$$\hat{c}^{\phi}_n \;=_d\; c^{\phi}_n, \quad n = 1, 2, \cdots. \tag{9}$$

Proof. Since ϕ_n depends on a_n but not the service times, it suffices to establish that for ϕ, the service times in any queue i, $1 \leq i \leq s$, in the system $\mathcal{S}(\pi, \rho)$ form a sequence of i.i.d. r.v.'s with distribution function $F(x)$.

The independence follows from the construction and the assumption that $\{u_k\}_{k=1}^{\infty}$ is a sequence of independent r.v.'s. Hence, it suffices to establish that each service time at queue i has distribution $F(x)$. Consider a service period under ϕ at server i that commences at time e_k for some $k \geq 1$. Consider how the service time, σ, for the customer commencing service at time e_k, is constructed.

We show by induction on m that the service time σ constructed in the time interval $[e_k, e_{k+m})$ has distribution $F(x)$. Clearly if $m = 1$, then $\sigma = F^{-1}(u_k)$ and, since u_k is uniformly distributed in $[0, 1)$, σ is distributed according to $F(x)$. Assume that the assertion holds for some $m \geq 1$. We consider the distribution of σ constructed in the time interval $[e_k, e_{k+m+1})$. If $x \leq e_{k+m} - e_k$, then clearly, by the inductive assumption, $\Pr[\sigma \leq x] = F(x)$. If, however, $x > e_{k+m} - e_k$, then

$$\Pr[\sigma \leq x] = \Pr[\sigma \leq e_{k+m} - e_k] + \Pr[e_{k+m} - e_k < \sigma \leq x]. \tag{10}$$

Using again the inductive assumption yields that

$$\Pr[\sigma \leq e_{k+m} - e_k] = F(e_{k+m} - e_k). \tag{11}$$

According to the construction of the service times, if σ exceeds $e_{k+m} - e_k$, then $F^{-1}_{e_{k+m}-e_k}(u_{k+m})$ will be given as the remaining service time. Therefore, cf. (3),

$$\Pr[e_{k+m} - e_k < \sigma \leq x] = (1 - F(e_{k+m} - e_k))F_{e_{k+m}-e_k}(x - e_{k+m} + e_k) = F(x) - F(e_{k+m} - e_k). \tag{12}$$

Combining relations (10)-(12) readily implies that when $x > e_{k+m} - e_k$, $\Pr[\sigma \leq x] = F(x)$. This completes the inductive proof. Hence the result. ■

We are now in a position to prove Theorem 1.

Proof of Theorem 1.

We focus on (1) first. Consider system $\mathcal{S}(\pi, \rho)$ with fixed arrival times and service parameters. It follows from property *(b)* of Lemma 2 that $\sum_{i=1}^{s} \widehat{N}_i^{\rho}(t) \leq \sum_{i=1}^{s} \widehat{N}_i^{\pi}(t)$. This coupled with property *(c)* of Lemma 2 is sufficient to ensure that

$$\widehat{N}^{\rho} \prec_w \widehat{N}^{\pi}(t), \quad t \geq 0.$$

Due to the characterizations of the weak majorization [15, pp. 108-109, Propositions B.1 and B.2], we have that for all increasing convex functions $g : I\!R \rightarrow I\!R$,

$$\sum_{i=1}^{s} g(N_i^{\rho}(t)) \leq \sum_{i=1}^{s} g(N_i^{\pi}(t)).$$

As a consequence of Lemma 3, we obtain that for all increasing convex functions $g : \mathbb{R} \to \mathbb{R}$,

$$E\left[\sum_{i=1}^{s} g(N_i^{\rho}(t)\right] \leq E\left[\sum_{i=1}^{s} g(N_i^{\pi}(t)\right]$$

provided the expectations exist, so that (1) holds.

$$\boldsymbol{N}^{\rho}(t) \leq_{E_3^{\uparrow}} \boldsymbol{N}^{\pi}(t), \quad t \geq 0.$$

We prove now (2). Consider system $\mathcal{S}(\pi, \rho)$. Fix the arrival times and the service time parameters $\{u_k\}$. Let customer n be the J_n^{π}-th departure from the s queues controlled by the routing policy π in system $\mathcal{S}(\pi, \rho)$. Let γ be the permutation on the set $\{J_n^{\pi} | 1 \leq n \leq m\}$ such that $\gamma(1) < \gamma(2) < \cdots < \gamma(m)$. It is easy to see that

$$\left(\hat{d}_1^{\pi} - a_1, \cdots, \hat{d}_m^{\pi} - a_m\right) \leq \left(\hat{d}_{\gamma(1)}^{\pi} - a_1, \cdots, \hat{d}_{\gamma(m)}^{\pi} - a_m\right).$$

It follows from standard interchange argument (cf. e.g., [1]) that

$$\left(\hat{d}_{\gamma(1)}^{\pi} - a_1, \cdots, \hat{d}_{\gamma(m)}^{\pi} - a_m\right) \prec_w \left(\hat{d}_{J_1^{\pi}}^{\pi} - a_1, \cdots, \hat{d}_{J_m^{\pi}}^{\pi} - a_m\right) = \left(\hat{c}_1^{\pi} - a_1, \cdots, \hat{c}_m^{\pi} - a_m\right).$$

Hence, by combining the above two inequalities, we obtain

$$\left(\hat{d}_1^{\pi} - a_1, \cdots, \hat{d}_m^{\pi} - a_m\right) \prec_w \left(\hat{c}_1^{\pi} - a_1, \cdots, \hat{c}_m^{\pi} - a_m\right). \tag{13}$$

On the other hand, it follows from Lemma 2 that

$$\left(\hat{c}_1^{\rho} - a_1, \cdots, \hat{c}_m^{\rho} - a_m\right) = \left(\hat{d}_1^{\rho} - a_1, \cdots, \hat{d}_m^{\rho} - a_m\right) \leq \left(\hat{d}_1^{\pi} - a_1, \cdots, \hat{d}_m^{\pi} - a_m\right).$$

This last relation together with (13) imply that

$$\left(\hat{c}_1^{\rho} - a_1, \cdots, \hat{c}_m^{\rho} - a_m\right) \prec_w \left(\hat{c}_1^{\pi} - a_1, \cdots, \hat{c}_m^{\pi} - a_m\right).$$

Thus, for all increasing convex functions $g : \mathbb{R} \to \mathbb{R}$,

$$\sum_{i=1}^{m} g(\hat{c}_i^{\rho} - a_i) \leq \sum_{i=1}^{m} g(\hat{c}_i^{\pi} - a_i). \tag{14}$$

Taking the expectation and using Lemma 3 on both sides of (14) entail that for all increasing convex functions $g : I\!R \to I\!R$,

$$E\left[\sum_{i=1}^{m} g(R_i^\rho)\right] \le E\left[\sum_{i=1}^{m} g(R_i^\pi)\right]$$

provided the expectations exist, so that (2) holds:

$$\boldsymbol{R}^\rho(m) \le_{E_3^\uparrow} \boldsymbol{R}^\pi(m).$$

$\blacksquare$

4 Remarks

Theorem 1 still holds when the initial workloads at the s servers are i.i.d. random variables and are independent of the arrival and service times. The proof can be carried out by constructing system $\mathcal{S}(\pi, \rho)$ in such a way that all the initial workloads are equal to $G^{-1}(u_0)$, where G is the distribution function of the workloads and u_0 is uniformly distributed in $[0, 1)$.

If the sequences of customer response times $\{R_n^\rho\}_n$ and $\{R_n^\pi\}_n$ are uniformly integrable and converge in distribution to the random variables R^ρ and R^π, respectively, then, it can be shown (cf. [1]) from (2) that the customer response time in steady state is minimized in the sense of increasing convex ordering by the round robin policy:

$$R^\rho \le_{icx} R^\pi.$$

It is interesting to conjecture that the optimality of the RR policy holds for more general service time distributions. Unfortunately, the arguments presented in this paper do not extend beyond the case of IFR distributions. The proof of Lemma 1 is based on this assumption. We have no counterexample to this conjecture and believe that it is true.

Acknowledgments: The authors are grateful to the referee for the useful comments on both the contents and the style of the paper.

References

[1] F. Baccelli, Z. Liu, D. Towsley, "Extremal Scheduling of Parallel Processing Systems with and without Real-Time Constraints." Rapport de Recherche INRIA 1113, 1989. To appear in *Journal of the ACM.*

[2] C. S. Chang, "A New Ordering for Stochastic Majorization: Theory and Applications", IBM Technical report, RC 16028, 1990.

[3] C. S. Chang, X. L. Chao, M. Pinedo, "A Note on Queues with Bernoulli Routing", *Proc. 29th Conf. on Decision and Control*, Hawaii, December 1990.

[4] D. J. Daley, "Certain Optimality Properties of the First Come First Served Discipline for $G/G/s$ Queues", *Stochastic Processes and their Applications*, **25**, pp. 301-308, 1987.

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

[6] S. G. Foss, "Approximation of Multichannel Queueing Systems" (in Russian), *Sibirski Mat. Zh.*, Vol. 21, pp. 132-140, 1980. (Transl.: *Siberian Math. J.*, **21**, pp. 851-857, 1980.)

[7] S. G. Foss, "Comparison of Servicing Strategies in Multichannel Queueing Systems" (in Russian), *Sibirski Math. Zh.*, Vol. 22, pp. 190-197, 1981. (Transl.: *Siberian Math. J.*, **22**, pp. 142-147, 1981.)

[8] L. Gün, A. Jean-Marie, "Parallel Queues with Resequencing", to appear in *J. ACM.*

[9] B. Hajek, "Extremal Splittings of Point Processes." *Math. Oper. Res.*, Vol. 10, No. 4, pp. 543-556, 1985.

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

[11] A. Jean-Marie, Z. Liu, "Stochastic Comparisons for Queueing Models via Random Sums and Intervals", *Advances in Applied Probabilities*, Vol. 24, pp. 960-985, 1992.

[12] J. F. C. Kingman, "Inequalities in the Theory of Queues," *J. Roy. Stat. Soc.*, ser. B, Vol. 32, pp. 102-110, 1970.

[13] Z. Liu, D. Towsley, "Effects of Service Disciplines in $G/G/s$ Queueing Systems", COINS Technical Report, TR 92-26, 1992. To appear in the *Annals of Operations Research*, special issue on Queueing Networks.

[14] Z. Liu, D. Towsley, "Stochastic Scheduling in In-Forest Networks". COINS Technical Report, TR 92-27, 1992. Submitted to *Advances in Applied Probabilities.*

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

[16] R. Menich, "Optimality of Shortest Queue Routing for Dependent Service Stations", *Proc. 26th Conf. on Decision and Control*, pp. 1069-1072, 1987.

[17] R. Menich, R. F. Serfozo, "Optimality of Routing and Servicing in Dependent Parallel Processing Systems", *Queueing Systems*, **9**, pp. 403-418, 1991.

[18] P.D. Sparaggis, C.G.Cassandras, D. Towsley, "On the Duality Between Routing and Scheduling Systems with Finite Buffer Space," to appear in *IEEE Transactions on Automatic Control.*

[19] P.D. Sparaggis, D. Towsley, C.G. Cassandras, "Extremal properties of the SNQ and the LNQ policies in finite capacity systems with state-dependent service rates," to appear in *Journal of Applied Probability.*

[20] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models.* English translation (D.J. Daley editor), J.Wiley and Sons, New York, 1983.

[21] D. Towsley, P.D. Sparaggis and C.G. Cassandras, 'Optimal routing and buffer allocation for a class of finite capacity queueing systems', to appear in *IEEE Trans. on Auto. Control.*

[22] D. Towsley, P.D. Sparaggis, "Optimal routing in systems with ILR service time distributions", submitted to *J. Appl. Prob.*, Feb. 1993.

[23] O. A. Vasicek, "An Inequality for the Variance of Waiting Time Under a General Queueing Discipline." *Operations Research*, **25**, pp. 879-884, 1977.

[24] J. Walrand, *An Introduction to Queueing Networks.* Prentice Hall, 1988.

[25] R. R. Weber, "On the optimal assignment of customers to parallel queues", *J. of Applied Prob.*, **15**, pp. 406-413, 1978.

[26] W. Whitt, "Deciding Which Queue to Join: Some Counterexamples." *Oper. Res.*, Vol. 34, No. 1, pp. 55-62, 1986.

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

[28] R. W. Wolff, "An Upper Bound for Multi-Channel Queues," *J. Appl. Prob.*, Vol. 14, pp. 884-888, 1977.

[29] R. W. Wolff, "Upper Bounds on Work in System for Multichannel Queues," *J. Appl. Prob.*, Vol. 24, pp. 547-551, 1987.