

# On the Optimality of FCFS for Networks of Multi-Server Queues

Ger Koole

*Vrije Universiteit*

*De Boelelaan 1081a, 1081 HV Amsterdam*

*The Netherlands*

Technical Report BS-R9235, CWI, Amsterdam, 1992

## Abstract

We consider multi-server queues in which arriving customers can be assigned to different servers. For three models the optimality of assigning customers to the server with smallest workload, FCFS, is shown. In the first two models isolated queues are studied, their difference being that in the first the cost functions are related to the workload, while in the second the departure processes are compared. The third model is concerned with networks of multi-server queues.

*1991 Mathematics Subject Classification:* 60K25, 90B22

*Keywords and Phrases:* FCFS, multi-server queues, optimal routing policies, networks of queues

*Note:* Supported by the European Grant BRA-QMIPS of CEC DG XIII

## 1. INTRODUCTION

In this paper we study assignment policies for multi-server queues. On arrival a customer has to be assigned to one of  $m$  servers, knowing the workload of each server, but not the service time of the arriving customer; the service times constitute a sequence of independent random variables with common distribution function  $P(\cdot)$ . Each server itself works in a FIFO fashion. An alternative way to look at the system is to consider it consisting of  $m$  parallel queues, each with a single server, in which each customer has to be assigned to a queue.

This system has been studied for two different types of objective functions, dealing with either the workload in the system or the departure process. For the first type of objective it was shown by Foss [4], Daley [2] and Wolff [16] (for a smaller class of policies) that sending each arriving customer to the server with the smallest amount of work minimizes the vector of server workloads in a Schur convex sense. This policy is equivalent to first-come first-serve (FCFS) or, when viewing the model as  $m$  parallel single-server queues, to the Smallest Workload Policy (SWP). This result holds for each point in time  $t$ , but not jointly over all  $t$ : it is not a pathwise result. In fact it is possible to construct a model for which it is impossible to combine the sample paths in such a way that the trajectories under FCFS have less workload than the corresponding trajectories under the

alternative policy for each  $t$ . This counterexample and a simple proof of the result of [2] and [4] using backward induction are discussed in section 2.

In section 3 we consider the departure process of a multi-server queue. Wolff [15] shows that FCFS gives a pathwise earlier departure process than any other assignment policy within the class of policies that do not depend on the workloads. Theorem 2 of Foss [3] states a more general result, namely that any increasing function of the departure times is minimized by FCFS, for all assignment policies. From this the pathwise optimality of FCFS follows directly. However, Foss [3] does not supply a full proof. We give one here using a simple coupling argument. We also prove the following monotonicity property of FCFS: when customers arrive earlier at a queue, they depart earlier from that queue.

In section 4 we use the monotonicity and the optimality of FCFS to prove that in a network of multi-server queues with static routing between the queues (that is, the decision on where to route the  $n$ th customer departing from station  $j$  is taken in advance and does not depend on the state of the system at the time of the transition) FCFS should be used in each station to get earlier customer streams throughout the network. This is a surprising result; for many models for which there is a simple policy for an isolated center, there are no general network results. For example, in the model where there is no information on the workloads, but just on the number of customers at the different servers, shortest queue routing is only optimal in centers without feedback to the network (Hordijk & Koole [6]); even for tandem models it can be shown that in general shortest queue routing is not optimal in all but the last center (Hordijk & Koole [5]). This is done by showing that shortest queue routing, in contrast with FCFS, is not monotone: earlier arrivals do not necessarily give earlier departures.

Many papers address optimization problems for  $G|GI|m$  queues. In the class of problems which contains the model studied here, each arriving customer has to be sent to a server at the moment of its arrival. A good way to think of these types of models is to consider each server as having its own queue. For different information structures the optimal policy has been obtained, often with additional constraints on the service times, like increasing hazard rates. If there is no information on the state of the queues, cyclic routing is optimal (proposition 8.3.4 in Walrand [14]). The case where the queue lengths are known has already been referred to: shortest queue routing is optimal (e.g., proposition 8.3.2 in [14]). The case where the workloads are available is the subject of the present paper. Also models with delayed queue length information have been studied (Kuri & Kumar [8], Koole [7]). Another class of problems consists of those in which the controller selects amongst the available customers those to work on (either preemptively or non-preemptively), without knowing their actual service times. Righter & Shanthikumar [12] and Liu & Towsley [10] are two recent references. When the service times are known to the controller or when the customers arrive in different classes we have (deterministic or stochastic) scheduling problems. Two general references are Lawler et al. [9] (deterministic scheduling) and Righter [11] (stochastic

scheduling).

## 2. MINIMIZING WORKLOADS

In this section we show that FCFS minimizes each weak Schur convex cost function stochastically at any time  $t$ , but first we formally introduce the model and our notation.

Customers arrive at times  $0 = t_1 \leq t_2 \leq \dots$ . (This can be seen as a realization of a general arrival process  $T$ ;  $t_1 = 0$  is taken merely for convenience.) Assume there are  $k$  arrivals in the interval  $[0, t)$ . It is convenient to number the interarrival times from  $t$  backwards:  $\tau_0 = t - t_k$ ,  $\tau_n = t_{k-n+1} - t_{k-n}$  for  $1 \leq n \leq k - 1$ .

Let the servers work at speed  $c$ , so that a server busy throughout an interarrival time of length  $\tau_n$  reduces a server's workload by  $u_n = c\tau_n$ . Let  $x = (x_1, \dots, x_m)$  denote a vector of workloads,  $x \in \mathbb{R}_+^m$ , and for  $y \in \mathbb{R}^m$  let  $y^+ = (y_1^+, \dots, y_m^+) \in \mathbb{R}_+^m$ . Define  $e = (1, \dots, 1)$  and  $e_j = (0, \dots, 0, 1, 0, \dots, 0)$ , with the 1 in the  $j$ th position.

Let  $v_x^0$  be the costs associated with reaching state  $x \in \mathbb{R}_+^m$  at time  $t$ . To find the policy which minimizes these costs we formulate the dynamic programming equation for our model. Let  $v_x^n$ ,  $n \geq 1$ , be the expected minimal costs starting at  $t_{k-n+1}$ , just before the arrival, with initial state  $x \in \mathbb{R}_+^m$ . Then

$$v_x^{n+1} = \min_j \left\{ \int_0^\infty v_{(x+se_j-u_n e)^+}^n dP(s) \right\} \quad (2.1)$$

for  $n = 0, \dots, k - 1$ . From the order of minimization and integration it follows that the decision is taken without knowing the actual service time of the arriving customer. The relation ((2.1)) preserves certain properties of the  $v^n$  as in the following lemma.

### 2.1. Lemma. If

$$\int v_{x+se_{j_1}}^n dP(s) \leq \int v_{x+se_{j_2}}^n dP(s) \text{ for } x_{j_1} \leq x_{j_2}, \quad (2.2)$$

$$v_x^n \leq v_{x+se_{j_1}}^n \quad \text{for } s \geq 0 \quad (2.3)$$

and

$$v_x^n = v_{x^*}^n \quad \text{for } x^* \text{ a permutation of } x \quad (2.4)$$

hold for  $n = 0$ , then they hold for all  $n$ .

The proof of lemma 2.1 starts by showing that

$$\int v_{(x+se_{j_1}-u_n e)^+}^n dP(s) \leq \int v_{(x+se_{j_2}-u_n e)^+}^n dP(s) \text{ for } x_{j_1} \leq x_{j_2} \quad (2.5)$$

follows from the inequalities. This inequality shows that assigning to the server with smallest workload is optimal. Thus the lemma gives conditions on  $v^0$ , the cost function, for FCFS to be optimal.

**Proof of lemma 2.1.** By induction. We show that (2.5) holds for all  $u = u_n$ . First assume that  $x_{j_1} - u \geq 0$ . This means that  $(x + se_j - ue)^+ = (x - ue)^+ + se_j$  for  $j = j_1$  and  $j = j_2$ . Then we have

$$\begin{aligned} \int v_{(x+se_{j_1}-ue)^+}^n dP(s) &= \int v_{(x-ue)^++se_{j_1}}^n dP(s) \leq \\ &\int v_{(x-ue)^++se_{j_2}}^n dP(s) = \int v_{(x+se_{j_2}-ue)^+}^n dP(s), \end{aligned}$$

where the inequality follows from ((2.2)).

Now assume that  $x_{j_1} - u < 0$ , but  $x_{j_2} - u \geq 0$ . By (2.3), monotonicity, we have  $v_{(x+se_{j_1}-ue)^+}^n \leq v_{(x-ue)^++se_{j_1}}^n$ . This gives

$$\begin{aligned} \int v_{(x+se_{j_1}-ue)^+}^n dP(s) &\leq \int v_{(x-ue)^++se_{j_1}}^n dP(s) \leq \\ &\int v_{(x-ue)^++se_{j_2}}^n dP(s) = \int v_{(x+se_{j_2}-ue)^+}^n dP(s). \end{aligned}$$

Here ((2.2)) is used to obtain the second inequality.

Finally assume that  $x_{j_2} - u < 0$ . We can rewrite  $(x + se_{j_2} - ue)^+$  as  $(x - ue)^+ + s^* e_{j_2}$  with  $s^* = (s - u + x_{j_2})^+$ . Note that  $s^* \leq s$ . Because  $(x + se_{j_1} - ue)^+ \leq (x - ue)^+ + s^* e_{j_1}$  we have, by (2.3),  $v_{(x+se_{j_1}-ue)^+}^n \leq v_{(x-ue)^++s^*e_{j_1}}^n$ . Thus

$$\begin{aligned} \int v_{(x+se_{j_1}-ue)^+}^n dP(s) &\leq \int v_{(x-ue)^++s^*e_{j_1}}^n dP(s) = \\ &\int v_{(x-ue)^++s^*e_{j_2}}^n dP(s) = \int v_{(x+se_{j_2}-ue)^+}^n dP(s). \end{aligned}$$

Having shown that assigning according to FCFS is optimal, the inequalities follow quite easily.

First we prove (2.2). Let  $x$  be such that  $x_{j_1} \leq x_{j_2}$  and take  $j^*$  such that  $x_{j^*} = \min_j \{(x + se_{j_2})_j\}$ . By (2.5) we see that  $j^*$  is the optimal assignment in  $x + se_{j_2}$ . Note that we can choose  $j^* \neq j_2$  and that  $j^*$  does not depend on  $s$ . If  $j^* = j_1$ , then

$$\begin{aligned} \int \min_j \left\{ \int v_{(x+se_{j_1}+ue_j-u_n e)^+}^n dP(u) \right\} dP(s) &\leq \int \int v_{(x+se_{j_1}+ue_{j_2}-u_n e)^+}^n dP(u) dP(s) = \\ &\int \min_j \left\{ \int v_{(x+se_{j_2}+ue_j-u_n e)^+}^n dP(u) \right\} dP(s) = \int v_{x+se_{j_2}}^{n+1} dP(s). \end{aligned}$$

If  $j^* \neq j_1$ , then

$$\begin{aligned} \int \min_j \left\{ \int v_{(x+se_{j_1}+ue_j-u_n e)^+}^n dP(u) \right\} dP(s) &\leq \int \int v_{(x+se_{j_1}+ue_{j^*}-u_n e)^+}^n dP(u) dP(s) \leq \\ &\int \int v_{(x+se_{j_2}+ue_{j^*}-u_n e)^+}^n dP(u) dP(s) = \int \min_j \left\{ \int v_{(x+se_{j_2}+ue_j-u_n e)^+}^n dP(u) \right\} dP(s), \end{aligned}$$

the second inequality by the optimality of the SWP as shown above.

Concerning (2.3), if  $j^*$  is the optimal action in  $x + se_{j_1}$ , we have

$$\min_j \left\{ \int v_{(x+ue_j-u_n e)^+}^n dP(u) \right\} \leq \int v_{(x+ue_{j^*}-u_n e)^+}^n dP(u) \leq \int v_{(x+se_{j_1}+ue_j-u_n e)^+}^n dP(u) = \min_j \left\{ \int v_{(x+se_{j_1}+ue_j-u_n e)^+}^n dP(u) \right\}.$$

The second inequality follows from ((2.3)).

Equation (2.4), symmetry, is trivial to prove.  $\square$

For the model with exponential service times and policies based on the numbers of customers instead of workloads similar inequalities as in lemma 2.1 exist (equations (4.3) to (4.5) in [6]).

Equation (2.2) without the integration, i.e.  $w_{x+se_{j_1}} \leq w_{x+se_{j_2}}$  for all  $s$ , is not true; this means that it is essential that the controller does not know the actual service times of the arriving customers. To construct an example illustrating this, take  $m = 2$ ,  $u_0 = 2$  and  $v_{(x_1, x_2)}^0 = x_1 + x_2$ , which indeed satisfies the conditions of lemma 2.1. Let the service time be equal to 2 a.s. Then it is easily seen that, if we take  $x = (0, 1)$ ,  $t = 1$ ,  $j_1 = 1$  and  $j_2 = 2$ , then  $v_{x+se_{j_1}}^1 = v_{(1,1)}^1 = 1 > 0 = v_{(0,2)}^1 = v_{x+se_{j_2}}^1$ .

By considering  $v^k$  we have the following.

**2.2. Theorem.** *FCFS minimizes the costs at  $t$  for all initial states and for all cost functions satisfying (2.2) to (2.4).*

In [6] it is shown that the class of cost functions satisfying the equations (4.3) to (4.5) in that paper is the class of weak Schur convex functions. Although the cost functions considered here are functions of  $\mathbb{R}_+^m$ , it is readily seen that again all Schur convex functions satisfy the inequalities. If we require the inequalities to hold for all service time distributions  $P$ , then the Schur convex functions are exactly the allowable cost functions. Examples of weak Schur convex functions are  $\max_j \{x_j\}$  and  $\sum_j x_j$ . If  $c_x$  is Schur convex, then so is  $\mathbb{I}_{\{c_x > s\}}$  for all  $s$ , meaning that each Schur convex cost function is not only minimized by FCFS in expectation, but also stochastically. Note that the statement in the penultimate paragraph of p. 304 in Daley [2], on the functions that respect weak majorization, is not correct: for example indicator functions of allowable cost functions are in general not convex.

We can generalize the model slightly by allowing all servers to go jointly on vacation or to have partial availability of all servers. This could be done by taking  $u_n = c_n$ , where  $c_n$  is the total availability of each server between  $t_{k-n}$  and  $t_{k-n+1}$ . Of course the servers should, at all times, all have the same availability.

For the model with exponential service times and decisions based on the numbers of customers at the servers it is well known that shortest queue routing is pathwise optimal (e.g. Walrand [14]). Here however we have the striking result that FCFS minimizes the total workload stochastically but

not pathwise. To construct a counterexample to the pathwise optimality, take a model with initial workload  $x = (1, 2)$  and speed  $c = 1$ . For the service time  $S$  we have  $\mathbb{P}(S = 1) = \mathbb{P}(S = 2) = \frac{1}{2}$ . The first customer arrives at  $t_1 = 0$ , the second at  $t_2 = 1$ . No more arrivals occur before 4, i.e.  $t_3 \geq 4$ . When we fix the policy used, there are 4 different realizations up to time 3, each with probability  $\frac{1}{4}$ . To get a pathwise ordering, we have to combine the realizations for FCFS and an arbitrary policy  $R$  such that FCFS is better for all  $t$ . Take  $R$  such that we start with assigning to the longest queue, but the second customer is assigned to the shortest. Let  $s_i$  ( $\tilde{s}_i$ ) denote the service time of the  $i$ th arriving customer in the model that uses FCFS ( $R$ ). At  $t = 1$  the amount of work is  $1 + s_1 + s_2$  ( $1 + \tilde{s}_1 + \tilde{s}_2$ ). Therefore we have to couple  $s_1 = s_2 = 1$  with  $\tilde{s}_1 = \tilde{s}_2 = 1$ . Now we show that if  $\tilde{s}_1 = 1$  and  $\tilde{s}_2 = 2$ , then there is no choice of  $s_1$  and  $s_2$  which is pathwise better. Take first  $s_1 = 1$  and  $s_2 = 2$ . Then, at  $t = 3$ , the system ruled by  $R$  is empty, but not the model under FCFS. For both eventualities with  $s_1 = 2$  we have that the amount of work just after the first arrival is larger under FCFS.

Related counterexamples can be found in Stoyan [13] and Asmussen [1]. Both show for a fixed coupling that FCFS is not better at all  $t$ : Stoyan [13] couples the service times under FCFS and an alternative policy in order of arrival (this is equivalent to  $s_1 = \tilde{s}_1$  and  $s_2 = \tilde{s}_2$  in our example), while Asmussen (in problem 1.1 of chapter 11 of [1]) couples the service times in the order in which they enter service (which is equivalent to  $s_1 = \tilde{s}_2$  and  $s_2 = \tilde{s}_1$ ).

### 3. MINIMIZING DEPARTURE TIMES

In this section we consider the departure processes of multi-server queues. To do so we need to be able to compare arrival and departure processes. We use the definitions given in Hordijk & Koole [5]. We regard an arrival process  $T = \{T_n, n \in \mathbb{N}\}$  as a sequence of arrival times, where  $T_n$  is the time of the  $n$ th arrival. (When there are fewer than  $k$  arrivals  $T_n = \infty$  for  $n \geq k$ .) For arrival processes  $T = \{T_n, n \in \mathbb{N}\}$  and  $\tilde{T} = \{\tilde{T}_n, n \in \mathbb{N}\}$  we say that  $T$  is pathwise earlier than  $\tilde{T}$  (written as  $T \leq_p \tilde{T}$ ) if they can be coupled such that, for coupled realizations  $t_1 \leq t_2 \leq \dots$  and  $\tilde{t}_1 \leq \tilde{t}_2 \leq \dots$  of  $T$  and  $\tilde{T}$ , we have  $t_n \leq \tilde{t}_n$  for all  $n$ . Note that we do not just couple the times of the  $n$ th arrival, but we assume that all arrival times are coupled jointly. We use a similar definition and notation for departure processes. The main purpose of this section is to show that earlier arrivals give earlier departures when comparing two centers in both of which FCFS is used, and that, for the same arrival process, FCFS gives earlier departures than an arbitrary (allowable) policy  $R$ . Throughout we assume that  $c = 1$ , i.e. the servers are working at unit speed.

Consider two multi-server queues, with arrival processes  $T$  and  $\tilde{T}$ , with  $T \leq_p \tilde{T}$ . Furthermore, these queues have initial work  $W = (W_1, \dots, W_m)$  and  $\tilde{W} = (\tilde{W}_1, \dots, \tilde{W}_m)$  which can be coupled such that for coupled realizations of the workloads  $w$  and  $\tilde{w}$  we have  $w_{\pi(j)} \leq \tilde{w}_j$  for all  $j$ , with  $\pi$  a permutation, and for the realizations of the departures  $u_n$  and  $\tilde{u}_n$  due to the initially available

customers we have  $u_n \leq \tilde{u}_n$  for all  $n$ . Thus the model with arrivals  $T$  has less initial work, and the initial customers leave earlier. Assume that in both queues FCFS is used. We have the following for the departure processes  $U$  and  $\tilde{U}$  (belonging to  $T$  and  $\tilde{T}$  respectively).

**3.1. Theorem.**  $U \leq_p \tilde{U}$ .

**Proof.** Take two coupled realizations of the arrival processes  $t_n$  and  $\tilde{t}_n$ . Take realizations of the initial work as described in the definition. We make the coupling complete by giving the  $n$ th arriving customer in both queues the same service time  $s_n$ . Let  $r_n$  be the moment at which the  $n$ th customer joins a server. We use the same recursion for  $r_n$  and  $u_n$  as used in Wolff [15]:

$$u_n = \text{nth order statistic of } \{r_i + s_i \mid i < n + m\}$$

$$r_n = \max\{t_n, u_{n-m}\}$$

If we assume  $w_1 \leq \dots \leq w_m$ , we take  $u_n = w_{m+n}$  if  $-m + 1 \leq n \leq 0$ . We have similar definitions for the model with later arrivals.

Now we show that the departures due to the customers not initially available are earlier for the model with earlier arrivals. By superposition of these departures with the departures of the customers initially available we prove our result.

We use induction on the number of departures. As induction basis we have  $r_n \leq \tilde{r}_n$  for  $n \leq m$ , because  $u_n \leq \tilde{u}_n$  for  $n \leq 0$ . Take  $n \geq m$ . Assume that  $r_k \leq \tilde{r}_k$  for  $1 \leq k \leq m$ . Then  $u_{k-m+1} \leq \tilde{u}_{k-m+1}$ , and thus  $r_{k+1} \leq \tilde{r}_{k+1}$ . Thus as a by-product of the induction we get  $u_n \leq \tilde{u}_n$  for all  $n$ , which gives the ordering of the departures.  $\square$

In the proof of the optimality of FCFS we take  $T = \tilde{T}$ , but with possibly different initial loads. For the network models of section 4 we need the result on different arrival processes.

We can generalize our result slightly by assuming that the arrival processes are ordered just up to  $t$ . In this case it is easy to see that the departure processes are also ordered up to  $t$ .

Now we consider two centers with the same arrival process  $T$ , but with different assignment policies. One model is governed by FCFS with departure process  $V$ ; the other has assignment policy  $R$  and departure process  $\tilde{V}$ . Of course  $R$  falls into the class of policies specified in the previous section; specifically it is not allowed to depend on the current and future service times.

The following theorem is referred to in the introduction for stating a result similar to that of theorem 2 of Foss [3].

**3.2. Theorem.**  $V \leq_p \tilde{V}$ .

**Proof.** We prove the result by fixing an arbitrary horizon  $t$ , and showing that the departures up to  $t$  are earlier under FCFS. Fix a realization of the arrival process and let  $t_1, \dots, t_k$  be the arrivals up

to  $t$ . We compare 2 policies  $R_1$  and  $R_2$ . Under  $R_1$  the service time of the  $n$ th arriving customer is  $s_n$ . Assume that it is pathwise optimal to use FCFS from arrival  $k^*$  onward to arrival  $k$ . Suppose that  $R_1$  uses FCFS from  $k^*$  on, but does not use FCFS at arrival  $k^* - 1$ . First we construct  $R_2$  such that  $R_2$  is pathwise better than  $R_1$  and uses FCFS at arrival  $k^* - 1$ . Using induction we then obtain that FCFS should also be used for arrival  $k^* - 1$  up to  $k$ .

Let us define  $R_2$ . Assume that  $R_1$  assigns the  $(k^* - 1)$ th customer to queue  $j_1$  (which is not the queue with the smallest workload), and the  $k^*$ th customer to queue  $j_2$  (by induction, according to FCFS). Now  $R_2$  assigns customer  $1, \dots, k^* - 2$  the same as  $R_1$ , assigns customer  $k^* - 1$  to  $j_2$  and  $k^*$  to  $j_1$ , and uses FCFS afterwards (which gives possibly different assignments than under  $R_1$ ). The coupling of the service times is taken differently for different customers and realizations. The  $k^* - 2$  first arriving customers under  $R_2$  have the same service times as the  $k^* - 2$  first arriving customers under  $R_1$ . Let  $w^n$  ( $\tilde{w}^n$ ) denote the workload at the  $n$ th arrival under  $R_1$  ( $R_2$ ). Note that  $w^{k^*-1} = \tilde{w}^{k^*-1}$ . Let  $\tau = t_{k^*} - t_{k^*-1}$ . If  $w_{j_1}^{k^*-1} \geq \tau$ , then we assign  $s_{k^*}$  at  $t_{k^*-1}$  to queue  $j_2$  and  $s_{k^*-1}$  at  $t_{k^*}$  to queue  $j_1$ . If  $w_{j_1}^{k^*-1} < \tau$ , then we assign in the same order as for  $R_1$ . Let  $w^*$  ( $\tilde{w}^*$ ) be the workload just after the  $k^*$ th arrival. Then it is easily seen that  $w^* \leq \tilde{w}^*$  (possibly after exchanging  $j_1$  and  $j_2$ ) and that departures occur earlier in  $w^*$ . Now we can use theorem 3.1 with initial load  $w^*$  and  $\tilde{w}^*$  to conclude the proof.  $\square$

**3.3. Remark.** Examining the proof, we see that the only customers that matter are customers  $k^* - 1$  and  $k^*$ , and they are coupled such that the customers being served first have the same service times: service times are “distributed” in order of commencement of service. When repeating this argument, resulting in the coupling of FCFS and an arbitrary policy  $R$ , we see that all service times are handed out in the order at which customers start service. The same coupling is clearly used in Wolff [15,16], but also, as here, implicitly in Foss [3], Daley [2] and in the proof of lemma 2.1.

**3.4. Remark.** At first sight this proof seems to work also for the workload model, yet this would contradict the counterexample of the previous section. However, problems arise when  $t_{k^*-1} < t < t_{k^*}$ ; in that case the coupling should always be in order of arrival. Thus, to show optimality of FCFS for cost functions related to workloads, the coupling has to depend on  $t$ . This idea is used in Wolff [16].

Instead of looking at the departure processes we can also consider the sojourn times of the customers. It is clear that the sojourn time of the  $n$ th arriving customer is not minimized by FCFS: if the  $n - 1$  customers arriving earlier do not join the server with the smallest workload, then the waiting and sojourn times of the  $n$ th arriving customer, who does join the server with the smallest workload, are minimized. Thus, to lower the sojourn time of the  $n$ th customer the sojourn times of previously arriving customers are increased. This motivates us to consider the summed sojourn times of the first  $n$  customers. To do so, we first look at the number of customers in the system

at  $t$ . As this number is equal to the number of arrivals by  $t$  minus the number of departures, we derive the following directly from theorem 3.2.

**3.5. Corollary.** *The number of customers in a multi-server queue is pathwise minimized by FCFS.*

Now we are ready to consider the sojourn times of the first  $n$  arrivals. First observe that the arrivals after the  $n$ th do not influence the sojourn times of the first  $n$ , as each server operates in a FIFO manner. Thus we can omit all arrivals after the  $n$ th. Now it is easily seen that the total sojourn time is equal to the integral of the number of customers over time. By dividing by  $n$ , and letting  $n \rightarrow \infty$ , we can also consider the average sojourn time.

**3.6. Corollary.** *The total sojourn time of the first  $n$  customers and the average sojourn time (if it exists) are stochastically minimized by FCFS.*

It follows directly from this corollary that the expected average sojourn time is also minimized by FCFS.

Instead of looking at sojourn times, we can also consider waiting times. From corollary 3.6 it does not follow that FCFS minimizes the total waiting time of the first  $n$  customers stochastically, although it follows that the total expected waiting time is minimized by FCFS. However, we can repeat the proof of theorem 3.2 with the summed waiting times as objective, from which we conclude that FCFS also minimizes the total waiting time of the first  $n$  customers stochastically.

#### 4. NETWORKS OF MULTI-SERVER QUEUES

The results of the previous section can be combined as follows. Suppose we have two queues with arrival processes  $T$  and  $\tilde{T}$ , of which  $T$  is earlier up to  $t$ , which are operated by FCFS and an arbitrary  $R$  respectively. Then, by first comparing two queues operated by FCFS with arrivals  $T$  and  $\tilde{T}$ , and then using the optimality of FCFS, we conclude that the departures from the first queue are earlier up to  $t$ ; the coupling used to derive this is in order of commencement of service (which is independent of  $t$ ).

In this section we consider a network of  $c$  queues, where routing between the queues is according to static rules. We call an assignment rule (for the departure process of queue  $i$ ) static if it is defined by a sequence of random variables  $\{\Pi_n, n \in \mathbb{N}\}$ , where  $\Pi_n = j$  corresponds to routing the  $n$ th departing customer (from queue  $i$ ) to queue  $j$ . The routing probabilities are stochastically independent of all queue lengths and arrival times, but need not be independent themselves. If all  $\Pi_n$  are equally distributed and independent, we have random routing (like in standard Jackson networks). Another example is cyclic assignment, by taking  $\mathbb{P}(\Pi_{n+1} = j + 1 \pmod{m} \mid \Pi_n = j) = 1$  for all  $n \geq 1$ , and  $\Pi_1$  arbitrary.

The queues themselves can be of different types. There can be multi-server queues such as we have studied in the previous sections, or queues of other types, as long as they satisfy the optimality/monotonicity property described for FCFS in the first paragraph of this section.

The model can be open, closed or a mixture of these. Let  $R$  be an arbitrary policy for the whole network, based on total information. Let  $T(i, j)$  be the departure process from queue  $i$  consisting of the customers routed to queue  $j$ , using FCFS (or, more generally, the locally optimal policy  $R^*$ ) in each center. The streams under the alternative policy  $R$  are denoted by  $\tilde{T}(i, j)$ . Outside arrivals are assumed to be coming from center 0. Note that  $R$  is allowed to depend on the state of the whole system, that is, it uses global information. The following theorem states that all arrival processes are earlier if the policy which is optimal in the case that only local information is available (like FCFS) is used in each queue.

**4.1. Theorem.**  $T(i, j) \leq_p \tilde{T}(i, j)$  for all  $i$  and  $j$ .

**Proof.** Due to the (possible) feedback in the network, arrival times depend on prior departure times. Therefore we cannot just consider the arrival and departure processes consecutively as we could have done for tandem systems. For general networks we have to use induction on the events in the whole system.

We couple the networks, one using FCFS ( $R^*$ ) and one using  $R$ , by constructing  $T^*(i, j)$  and  $\tilde{T}^*(i, j)$  with  $T^*(i, j) \stackrel{d}{=} T(i, j)$  and  $\tilde{T}^*(i, j) \stackrel{d}{=} \tilde{T}(i, j)$  for all  $i$  and  $j$ . The routing is coupled by letting the  $n$ th customer that leaves center  $i$  go to the same center in both networks. Note that, by taking  $i = 0$ , we have  $T^*(0, j) = \tilde{T}^*(0, j)$ . The service times are coupled for each queue separately, such that the departures are earlier under FCFS. Now consider a realization.

Events in the networks with streams  $T^*$  and  $\tilde{T}^*$  occur at points  $t_1 < t_2 < \dots$  and  $\tilde{t}_1 < \tilde{t}_2 < \dots$ . Each event consists of a transition of a customer from one center to another. Transitions from center  $i$  to center  $j$  occur at  $t_1(i, j) < t_2(i, j) < \dots$  and  $\tilde{t}_1(i, j) < \tilde{t}_2(i, j) < \dots$ . (If 2 or more events occur at the same time, we assume that they are logically ordered. For example, if a customer arrives at a center, receives 0 processing time and leaves again, we assume that the arrival occurs before the departure.)

We use the fact that if the arrivals up to  $T$  at a certain center are earlier in the FCFS model, then the departures up to  $T$  are earlier also. The proof uses induction on the number of events in the network operated by  $R$ . Choose  $n^*$ . Define  $n_{ij}^*$  as follows:  $\tilde{t}_{n_{ij}^*}(i, j) \leq \tilde{t}_{n^*} < \tilde{t}_{n_{ij}^*+1}(i, j)$ . Suppose

$$t_l(i, j) \leq \tilde{t}_l(i, j) \text{ for all } l = 1, \dots, n_{ij}^*, i \text{ and } j.$$

Consider transition  $n^* + 1$  in the network operated by  $R$ . Suppose that a customer moves from center  $i^*$  to center  $j^*$  at this transition. Consider center  $i^*$ . By the induction hypothesis for  $j = i^*$ , the arrivals at  $i^*$  before  $\tilde{t}_{n^*}$  are earlier under FCFS. Because there are no arrivals at center  $i^*$

between  $\tilde{t}_{n^*}$  and  $\tilde{t}_{n^*+1}$  in the network operated by  $R$ , also the arrivals before  $\tilde{t}_{n^*+1}$  are earlier under FCFS. By the optimality and monotonicity of FCFS, the departures are also earlier, and thus  $t_{n_{i^*j^*}^*+1} \leq \tilde{t}_{n_{i^*j^*}^*+1}$ , completing the induction step.  $\square$

An example of the type of center which also has the monotonicity/optimality property is the single-server queue studied by Righter & Shanthikumar [12]. They show that in the case of service time distributions with increasing likelihood ratios, the departures are earlier if the customers are served non-preemptively.

Besides controllable centers we can also add queues which have just a single policy, but for which the monotonicity property holds. Examples of these are  $G|G|\infty$  queues and  $G|G|1$  queues with FIFO discipline.

Our results can be summarized as follows.

**4.2. Corollary.** *In a closed network, FCFS ( $R^*$ ) maximizes the throughput at each queue. In an open network, FCFS ( $R^*$ ) minimizes the number of customers in the system.*

For their model Righter & Shanthikumar [12] formulate a similar network result.

As for the isolated centers of the previous section we can look at waiting times. Unfortunately, the situation is more complex for networks. Due to the different routes customers can choose in the network, a customer can influence the waiting times of customers who had arrived earlier. Thus we cannot restrict the arrival process to the first  $n$  customers as we did for the isolated queue. However, by looking at the numbers of customers in the system for all  $t$ , we have the following corollary.

**4.3. Corollary.** *In an open network, FCFS minimizes the expected average sojourn and waiting times of the customers.*

**Acknowledgment.** I like to thank Serguei Foss for pointing me to his work on this subject and for the interesting discussion that followed, and Onno Boxma for many valuable comments.

## REFERENCES

- [1] S. Asmussen. *Applied Probability and Queues*. Wiley, 1987.
- [2] D.J. Daley. Certain optimality properties of the first-come first-served discipline for  $G|G|s$  queues. *Stochastic Processes and their Applications*, 25:301–308, 1987.
- [3] S.G. Foss. Approximation of multichannel queueing systems. *Siberian Mathematical Journal*, 21:851–857, 1981.
- [4] S.G. Foss. *Extremal Problems in Queueing Theory*. PhD thesis, Novosibirsk State University, 1982. In Russian.
- [5] A. Hordijk and G.M. Koole. On the shortest queue policy for the tandem parallel queue. *Probability in the Engineering and Informational Sciences*, 6:63–79, 1992.
- [6] A. Hordijk and G.M. Koole. On the assignment of customers to parallel queues. *Probability in the Engineering and Informational Sciences*, 6:495–511, 1992.
- [7] G.M. Koole. Optimal repairman assignment in two symmetric maintenance models. *European Journal of Operations Research*, 82:295–301, 1995.
- [8] J. Kuri and A. Kumar. Optimal control of arrivals to queues with delayed queue length information. In *Proceedings of the 31th IEEE Conference on Decision and Control*, 1992.
- [9] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys. Sequencing and scheduling: Algorithms and complexity. In S.C. Graves, A.H.G. Rinnooy Kan, and P. Zipkin, editors, *Handbooks in Operations Research and Management Science, vol. 4: Logistics of Production and Inventory*. North-Holland, 1993.
- [10] Z. Liu and D. Towsley. Effects of service disciplines in  $G|GI|s$  queueing systems. *Annals of Operations Research*, 48:401–429, 1994.
- [11] R. Righter. Scheduling. In M. Shaked and J.G. Shanthikumar, editors, *Stochastic Orders and their Applications*, pages 381–432. Academic Press, 1994.
- [12] R. Righter and J.G. Shanthikumar. Extremal properties of the FIFO discipline in queueing networks. *Journal of Applied Probability*, 29:967–978, 1992.
- [13] D. Stoyan. A critical remark on a system approximation in queueing theory. *Mathematische Operationsforschung und Statistik*, 7:953–956, 1976.
- [14] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, 1988.
- [15] R.W. Wolff. An upper bound for multi-channel queues. *Journal of Applied Probability*, 14:884–888, 1977.
- [16] R.W. Wolff. Upper bounds on work in system for multichannel queues. *Journal of Applied Probability*, 24:547–551, 1987.