Effects of Service Disciplines in $G/GI/s$ Queueing Systems*

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

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

March 1992
Revised December 1992.

Abstract

Transient extremal properties of some service disciplines are established in the $G/GI/s$ queueing system for the minimization and maximization of the expectations of the Schur convex functions, convex symmetric functions and the sums of convex functions of the waiting times, response times, lag times and latenesses. When resequencing is required in the system, the FCFS and LCFS disciplines are shown to minimize and maximize, respectively, the expectations of any increasing functions of the end-to-end delays. All of these results are presented in terms of stochastic orderings. The paper concludes with extensions of the results to the stationary regime and to tandem as well as general queueing networks.

Keywords : Queueing System, Service Discipline, Stochastic Ordering, Sample Path Analysis, Lateness, End-to-End Delay.

*This work was supported in part by the National Science Foundation under grant ASC 88-8802764.
†The work of this author was also partially supported by CEC DG-XIII under the ESPRIT-BRA grant QMIPS

1 Introduction

Queueing models with multiple servers are frequently used in the performance analysis of multiprocessor systems (see e.g. [17]) and communication networks (see e.g. [14]), where the servers represent the processors or communication channels. In this paper, we analyze the effects of different service disciplines in such queueing systems. We will first focus on a simple $G/GI/s$ queueing model consisting of a single queue and s servers. The extension to queueing networks with multi-server queues will be discussed at the end of the paper.

In the $G/GI/s$ queue under consideration, the arrival times of the customers are arbitrary, whereas the service times are independent and identically distributed (i.i.d.) random variables (r.v.'s) independent of the arrival times. A customer can be served by any of the s servers. These servers are identical and have the same service rate, say 1.

When the service disciplines are non-preemptive and use only the information on the distribution of service times, it has been established by various authors (see Kingman [13], Vasicek [22], Foss [9, 10], Wolff [24, 25] and Daley [7]) that the First Come First Serve (FCFS) discipline minimizes the stationary waiting times in the sense of convex ordering. (Note that the transient results in Foss [9, 10], Wolff [24, 25] and Daley [7] are for the (e.g. Kiefer-Wolfowitz) workload vector).

When preemption is allowed and there is a single server $s = 1$, Shantikumar and Sumita [19] showed that if the service times have an Erlang distribution, then the FCFS discipline minimize the stationary waiting times in the sense of increasing convex ordering. This last result was generalized by Hirayama and Kijima [11], and Chang and Yao [6] to the case when the service time distribution is of Increasing Failure Rate (IFR) type.

When due dates (also called soft real-time constraints or soft deadlines) are associated with the customers, Pinedo [18] analyzed the expected weighted number of late jobs. Baccelli, Liu and Towsley [2] obtained extremal service disciplines, within a queueing network of single-server queues, for the vector of transient customer latenesses in the sense of Schur convex ordering.

The resequencing problem in communication networks is often analyzed using the multi-server queueing model (cf. Kleinrock et al. [14], Baccelli et al. [1, 3]). Whitt [23] analyzed the number of customers overtaken by an arbitrary customer for $GI/M/s$ and $M/GI/s$ models with FCFS service discipline. Iliadis and Lien [12] studied the resequencing delay for two heterogeneous servers under a threshold-type scheduling.

In this paper, we compare different service disciplines in the $G/GI/s$ queueing model with delay dependent customer behavior. More specifically, we analyze two cases:

- Each customer carries a due date with it. The performance metrics under consideration are the lag time (i.e., the difference between the service beginning time and the due date), and the lateness (i.e., the difference between the service completion time and the due date). When the

customer due dates are identical to the customer arrival times, the lag time and the lateness correspond to the customer waiting time and response time, respectively.

- There is a resequencing buffer. Every customer enters the buffer after being serviced and leaves when all of the previous arrived (to the queueing system) customers have left the resequencing buffer. The performance measure of interest is the end-to-end delay defined as the difference between the time when the customer leaves the resequencing buffer and its arrival time.

For the first case, we devise two new service disciplines, referred to as Stochastically Smallest Due Date (SSDD) and Stochastically Largest Due Date (SLDD), which are defined to be such that, as soon as there is an available server, the customer waiting in the queue with the stochastically smallest and largest due dates, respectively, is assigned to the server. Such disciplines include, as a special case, the Smallest Due Date (SDD) and Largest Due Date (LDD) disciplines. When the due dates are set to the arrival times, SSDD and SLDD coincide with FCFS and LCFS disciplines, respectively.

In accordance with the assumptions on the due dates and service times, these disciplines are shown to be extremal for the customer lag times and the latenesses in the sense of the E_1 ordering (i.e. Schur convex ordering), E_2 ordering (convex symmetric ordering) and E_3 ordering. All of these comparison results on the vector of transient performance metrics imply the convex ordering on the corresponding stationary performance metrics.

For the second case, we prove that FCFS and LCFS disciplines bound from below and above, respectively, the end-to-end delays in the sense of strong stochastic ordering. Owing to the fact that FCFS and LCFS disciplines also bound the response times in the sense of convex ordering, our results are consistent with Whitt's conjecture [23, Conjecture 1.3], which says that in an open Jackson network, the total sojourn time increases in the sense of convex ordering as customer overtaking increases.

Finally, the paper concludes with results for queueing networks. These include results for tandem queueing networks whose constituents are either $\cdot/D/s$ or $\cdot/M/s$ queues. In the latter case, the scheduling policies can be preemptive resume. Additional results for other classes of queuing networks are also given.

The paper is organized as follows. In the next section, we define in a more precise way the model, as well as the notation and assumptions. We also present some preliminaries on the stochastic orderings. In Section 3, we analyze the extremal properties within the class of non-idling and non-preemptive disciplines. In Section 4, we generalize the results to the class of non-idling preemptive disciplines under the assumption that the service times are exponentially distributed. In Section 5, we prove the optimality of the FCFS discipline within the class of idling (and possibly preemptive) disciplines. In Section 6, we extend the extremal properties to the stationary regime and to other performance metrics. These results are further extended to some queueing networks with multi-server queues.

2 Notation and Assumptions

2.1 Model Description

The queueing system under consideration is a $G/GI/s$ model, i.e., there are $s \geq 1$ servers associated with a single waiting queue and a resequencing buffer, both of infinite capacity. The servers are identical and have the same speed, say 1. When a customer enters the system, it waits for service in the waiting queue. After having been served by one of the s servers, the customer enters the resequencing buffer. A customer, say n, can leave the resequencing buffer (and, thus the system), if and only if all the customers $1, 2, \cdots, n-1$ have already left this buffer.

Let $N \geq 1$ be the arbitrarily fixed number of arrivals. Denote by a_n and σ_n the arrival time and the service time of customer n, respectively, $1 \leq n \leq N$, with $a_1 = 0 \leq a_2 \leq \cdots \leq a_n \leq \cdots \leq a_N$. The sequence of service times $\mathcal{S} = \{\sigma_n\}_{n=1}^N$ consists of i.i.d. random variables. The sequence of arrival times $\mathcal{A} = \{a_n\}_{n=1}^N$ is independent of the service times, but is otherwise arbitrary. In particular, it can be a deterministic sequence.

We associate with customer n, $1 \leq n \leq N$, a due date, denoted by d_n. Let $r_n = d_n - a_n$ be the relative due date of customer n. Both d_n and r_n are (not necessarily positive) real numbers. We assume that the sequence of due dates $\mathcal{D} = \{d_n\}_{n=1}^N$ is independent of the service times $\mathcal{S}$.

2.2 Service Disciplines

A service discipline decides the time at which a particular customer is to be served. The service discipline is called non-preemptive if the service of any customer cannot be stopped unless its service is finished. The discipline is preemptive (resume) if it is preemptive and if the service is resumed at the point where it was preempted. The discipline is called non-idling or work conserving if no server is allowed to stay idle whenever there is a customer waiting in the queue.

Throughout this paper we assume that the service disciplines cannot use the information on the exact values of the service times, but only the distribution of the service times. This assumption implies that the service disciplines like Shortest Remaining Processing Time are not under consideration. We also assume that the service disciplines are not anticipative in the sense that a decision can never use information on future arrivals.

Denote by Ψ the class of (possibly idling and/or preemptive) disciplines fulfilling the above assumptions, $\Psi_{np} \subset \Psi$ the class of non-preemptive (possibly idling) disciplines, and $\Psi_{ni} \subset \Psi$ the class of non-idling (possibly preemptive) disciplines, and finally, $\Psi_0 = \Psi_{np} \bigcap \Psi_{ni} \subset \Psi$ the class of non-idling and non-preemptive disciplines.

Among the well-known extremal disciplines, there are First Come First Serve (FCFS) and Last Come First Serve (LCFS) disciplines, which are defined to assign, as soon as possible, the first

and the last arrived customers to available servers, respectively. Note that here FCFS and LCFS disciplines are defined to be non-idling.

When the customers are associated with known due dates, there exist the Smallest Due Date (SDD) and the Largest Due Date (LDD) disciplines, which assign customers according to their due dates.

When the due dates are comparable in the strong stochastic ordering sense $\leq_{st}$ (see the definition below), which is the case when they are known or when they are unknown but the relative due dates are i.i.d. random variables, we define the disciplines Stochastically Smallest Due Date (SSDD) and Stochastically Largest Due Date (SLDD) to be such that as soon as there is an available server, the customer waiting in the queue with the stochastically smallest and largest due dates, respectively, is assigned to the server. Again, by definition, the SSDD and SLDD disciplines are non-idling.

Observe that when the relative due dates are i.i.d. random variables independent of the arrival and service times and are unknown a priori, then the SSDD and SLDD disciplines coincide with the FCFS and LCFS disciplines, respectively. When the due date are known, the SSDD and SLDD disciplines coincide with the SDD and LDD disciplines, respectively.

For the Preemptive (P) version of these disciplines, we will use the notation PLCFS, PSDD, PLDD, PSSDD, PSLDD, etc.

Within the general class of idling preemptive disciplines, we define the class of FCFS disciplines, denoted by Π_{FCFS}, to be such that every discipline in Π_{FCFS} assigns, when it decides to assign a customer, the first arrived customer among the waiting customers to an available server. It is clear that FCFS $\in \Pi_{FCFS}$.

2.3 Performance Metrics

Let $\pi \in \Psi$ be an arbitrary service discipline, π_n be the identity of the n-th assigned customer, $1 \leq n \leq N$. Denote by $b_n(\pi)$ (resp. $c_n(\pi)$) the random variable (in $I\!R^+$) of the service beginning (resp. completion) time of customer n. Denote by $W_n(\pi)$, $R_n(\pi)$, $l_n(\pi)$ and $L_n(\pi)$ the waiting time, the response time, the lag time and the lateness of customer n under $\pi \in \Psi$, respectively, defined as

$$W_n(\pi) = b_n(\pi) - a_n, \tag{2.1}$$

$$R_n(\pi) = c_n(\pi) - a_n, \tag{2.2}$$

$$l_n(\pi) = b_n(\pi) - d_n, \tag{2.3}$$

$$L_n(\pi) = c_n(\pi) - d_n. \tag{2.4}$$

Let

$$\boldsymbol{W}(\pi) = (W_1(\pi), \cdots, W_N(\pi)),$$

$$\boldsymbol{R}(\pi) = (R_1(\pi), \cdots, R_N(\pi)),$$

$$\begin{aligned} \boldsymbol{l}(\pi) &= (l_1(\pi), \cdots, l_N(\pi)), \\ \boldsymbol{L}(\pi) &= (L_1(\pi), \cdots, L_N(\pi)). \end{aligned}$$

Denote by $q_n(\pi)$ and $D_n(\pi)$ the departure time and the end-to-end delay, respectively, of customer n under $\pi \in \Psi$, i.e.,

$$q_n(\pi) = \max(c_n(\pi), q_{n-1}(\pi)) = \max_{1 \le l \le n} c_l(\pi), \tag{2.5}$$

$$D_n(\pi) = q_n(\pi) - a_n = \max_{1 \le l \le n} c_l(\pi) - a_n. \tag{2.6}$$

The purpose of this paper is to find extremal disciplines that minimize or maximize these performance measures in some stochastic semi-partial ordering sense. Hence, we define the partial orderings of interest to us in the remainder of this section.

2.4 Stochastic Orderings

Let $\boldsymbol{X}, \boldsymbol{Y} \in I\!R^n$ be two random vectors.

Definition 2.1 *The random vector $\boldsymbol{X}$ is stochastically less than the random vector $\boldsymbol{Y}$ in the sense of strong stochastic ordering ($\boldsymbol{X} \le_{st} \boldsymbol{Y}$), convex ordering ($\boldsymbol{X} \le_{cx} \boldsymbol{Y}$), and increasing convex ordering ($\boldsymbol{X} \le_{icx} \boldsymbol{Y}$), respectively, if*

$$E[f(\boldsymbol{X})] \le E[f(\boldsymbol{Y})], \quad \forall \text{ increasing } f : I\!R^n \to I\!R, \tag{2.7}$$

$$E[f(\boldsymbol{X})] \le E[f(\boldsymbol{Y})], \quad \forall \text{ convex } f : I\!R^n \to I\!R, \tag{2.8}$$

$$E[f(\boldsymbol{X})] \le E[f(\boldsymbol{Y})], \quad \forall \text{ increasing and convex } f : I\!R^n \to I\!R, \tag{2.9}$$

respectively, provided the expectations exist.

The reader is referred to [20] for properties concerning these orderings. In what follows, $=_{st}$ denotes equality in distribution. The following lemma is due to Strassen [21]:

Lemma 2.1 *Two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy $\boldsymbol{X} \le_{st} \boldsymbol{Y}$ if and only if there exist two random vectors $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ defined on a common probability space such that $\boldsymbol{X} =_{st} \widetilde{\boldsymbol{X}}$, $\boldsymbol{Y} =_{st} \widetilde{\boldsymbol{Y}}$, and $\widetilde{\boldsymbol{X}} \le \widetilde{\boldsymbol{Y}}$ componentwise almost surely (a.s.).*

Define now the notion of majorization. Let $x, y \in I\!R^n$ be two real vectors.

Definition 2.2 *Vector x is said to be majorized by vector y (written $x \prec y$) iff*

$$\sum_{i=1}^{k} \hat{x}_i \leq \sum_{i=1}^{k} \hat{y}_i, \quad k = 1, \cdots, n-1 \tag{2.10}$$

$$\sum_{i=1}^{n} \hat{x}_i = \sum_{i=1}^{n} \hat{y}_i, \tag{2.11}$$

where the notation $\hat{x}_i$ is taken to be the i-th largest element of x. If

$$\sum_{i=1}^{k} \hat{x}_i \leq \sum_{i=1}^{k} \hat{y}_i, \quad k = 1, \cdots, n \tag{2.12}$$

then vector x is said to be weakly majorized by vector y (written $x \prec_w y$)

Definition 2.3 *A function $f : I\!R^n \to I\!R$ is Schur-convex if for all $x, y \in I\!R^n$,*

$$x \prec y \quad \Rightarrow f(x) \leq f(y).$$

We define the following classes of functions

- $\mathcal{C}_1$ ($\mathcal{C}_1^{\uparrow}$) - the class of (increasing) Schur-convex functions,
- $\mathcal{C}_2$ ($\mathcal{C}_2^{\uparrow}$) - the class of (increasing) symmetric and convex functions,
- $\mathcal{C}_3$ ($\mathcal{C}_3^{\uparrow}$) - the class of functions of the form $f(\boldsymbol{X}) = \sum_{i=1}^{n} f(x_i)$ where f is (increasing) convex.

Definition 2.4 *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two random vectors in $I\!R^n$. We define the following stochastic orderings between these r.v.'s*

$$\boldsymbol{X} \leq_{E_i} \boldsymbol{Y}, \quad \text{if} \quad E[f(\boldsymbol{X})] \leq E[f(\boldsymbol{Y})], \quad \forall f \in \mathcal{C}_i,\ i = 1, 2, 3.$$

and

$$\boldsymbol{X} \leq_{E_i^{\uparrow}} \boldsymbol{Y}, \quad \text{if} \quad E[f(\boldsymbol{X})] \leq E[f(\boldsymbol{Y})], \quad \forall f \in \mathcal{C}_i^{\uparrow},\ i = 1, 2, 3.$$

Various properties concerning these orderings can be found in [16]. According to Proposition C.2 of [16, p. 67], any function $f : I\!R^n \to I\!R$ is Schur-convex if it is symmetric and convex. Therefore,

Proposition 2.1 *The following implication relations between the stochastic orderings hold:*

$$\begin{array}{ccccccc} & & \leq_{E_1} & \Rightarrow & \leq_{E_2} & \Rightarrow & \leq_{E_3} \\ & & \Downarrow & & \Downarrow & & \Downarrow \\ \leq_{st} & \Rightarrow & \leq_{E_1^{\uparrow}} & \Rightarrow & \leq_{E_2^{\uparrow}} & \Rightarrow & \leq_{E_3^{\uparrow}} \end{array}$$

The relations still hold when $\leq_{E_1}$ and $\leq_{E_1^{\uparrow}}$ are replaced by $\leq_{cx}$ and $\leq_{icx}$, respectively.

The following lemma is another application of Strassen's Theorem (cf. [21]) to the semi-partial orderings $\prec$ and $\prec_w$. A proof was provided in [16, Theorem B.1, p. 483].

Lemma 2.2 *Two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ satisfy $\boldsymbol{X} \leq_{E_1} \boldsymbol{Y}$ ($\boldsymbol{X} \leq_{E_1^{\uparrow}} \boldsymbol{Y}$, respectively) if and only if there exist two random variables $\widetilde{\boldsymbol{X}}$ and $\widetilde{\boldsymbol{Y}}$ defined on a common probability space such that $\boldsymbol{X} =_{st} \widetilde{\boldsymbol{X}}$, $\boldsymbol{Y} =_{st} \widetilde{\boldsymbol{Y}}$, and $\widetilde{\boldsymbol{X}} \prec \widetilde{\boldsymbol{Y}}$ ($\widetilde{\boldsymbol{X}} \prec_w \widetilde{\boldsymbol{Y}}$, respectively) a.s.*

3 Comparisons within the Class of Non-Idling and Non-Preemptive Service Disciplines

In this section, we focus on the class of non-idling and non-preemptive disciplines. The comparison of other service disciplines will be discussed in the following sections.

3.1 Lag Times and Waiting Times

Theorem 3.1 *Assume that the due dates are known. Then the SDD (resp. LDD) discipline minimizes (resp. maximizes) the vector of lag times within the class of non-preemptive non-idling disciplines Ψ_0 in the sense of $\leq_{E_1}$ ordering:*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{l}(SDD) \leq_{E_1} \boldsymbol{l}(\pi) \leq_{E_1} \boldsymbol{l}(LDD). \tag{3.1}$$

Proof. Consider an arbitrary discipline $\pi \in \Psi_0$. Let the sequence of arrival times $\mathcal{A}$ be arbitrarily fixed. Let m, $1 \leq m \leq N-1$, be the first time that π does not follow the SDD rule, i.e., $\pi_i = SDD_i$, $1 \leq i \leq m-1$. Suppose that at the m-th assignment, π selects customer j, whereas customer k would be chosen for service if the SDD rule were applied, i.e.,

$$d_j > d_k, \tag{3.2}$$

and $\pi_m = j$, $SDD_m = k$. Suppose further that $\pi_n = k$, namely, customer k is the n-th that starts service under π, $m < n \leq N$, which implies that

$$b_j(\pi) \leq b_k(\pi). \tag{3.3}$$

Construct now a non-idling discipline π' which differs from π only in the customer selections of j and k, viz.,

$$\pi'_m = \pi_n, \qquad \pi'_n = \pi_m, \qquad \pi'_i = \pi_i, \quad i \neq m, \ \ i \neq n, \ \ 1 \leq i \leq N.$$

The discipline π' operates on the sequence of service times $\mathcal{S}' = \{\sigma'_i\}_{i=1}^N$, with

$$\sigma'_j = \sigma_k, \qquad \sigma'_k = \sigma_j, \qquad \sigma'_i = \sigma_i, \quad i \neq j, \ \ i \neq k, \ \ 1 \leq i \leq N.$$

It is verified that $\mathcal{S}'$ is equivalent to $\mathcal{S}$ in law, and independent of the arrival and due dates (see [15] for a formal proof). One can check that

$$b_j(\pi') = b_k(\pi), \qquad b_k(\pi') = b_j(\pi), \qquad b_i(\pi') = b_i(\pi), \quad i \neq j, \quad i \neq k, \quad 1 \leq i \leq N, \tag{3.4}$$

so that π' is feasible.

It follows from the relations (3.2), (3.3) and (3.4) that

$$\boldsymbol{l}(\pi') \prec \boldsymbol{l}(\pi).$$

Since $\mathcal{S}'$ is identical to $\mathcal{S}$ in law, we have, conditioned on the arrival times, that for all Schur-convex functions $f : I\!R^N \to I\!R$,

$$E[f(\boldsymbol{l}(\pi'))|\mathcal{A}] \leq E[f(\boldsymbol{l}(\pi))|\mathcal{A}].$$

Unconditioning with respect to the arrival times in the above inequality yields

$$E[f(\boldsymbol{l}(\pi'))] \leq E[f(\boldsymbol{l}(\pi))]. \tag{3.5}$$

Consider now the discipline π'. If it follows the SDD rule everywhere, one gets the desired result. Otherwise, suppose $m_1 > m$ is the first time that it violates the SDD rule: $\pi'_{m_1} \neq SDD_{m_1}$. One can construct a discipline π'' in a similar way such that π'' follows the SDD rule until m_1 and that for any Schur-convex function $f : I\!R^N \to I\!R$,

$$E[f(\boldsymbol{l}(\pi''))] \leq E[f(\boldsymbol{l}(\pi'))].$$

After repeating this process for at most N times, one finally obtains the SDD discipline for which the relation

$$E[f(\boldsymbol{l}(SDD))] \leq E[f(\boldsymbol{l}(\pi))],$$

holds, provided the function $f : I\!R^N \to I\!R$ is Schur-convex. This completes the proof of the first inequality in (3.1).

In an analogous way, one can prove that for any Schur-convex function $f : I\!R^N \to I\!R$, the LDD discipline satisfies the relation

$$E[f(\boldsymbol{l}(\pi))] \leq E[f(\boldsymbol{l}(LDD))].$$

■

In the above theorem, if we set the due dates to be the arrival times, the SDD and LDD disciplines coincide with the FCFS and LCFS disciplines, respectively, and the lag times represent the waiting times. Thus, as an immediate consequence of the above theorem, we obtain

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{W}(FCFS) \leq_{E_1} \boldsymbol{W}(\pi) \leq_{E_1} \boldsymbol{W}(LCFS). \tag{3.6}$$

Remark: In [22], a weaker result was obtained:

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{W}(FCFS) \leq_{E_3} \boldsymbol{W}(\pi) \leq_{E_3} \boldsymbol{W}(LCFS).$$

Corollary 3.1 *Assume now that the due dates are unknown a priori. Assume further that the relative due dates are i.i.d. random variables that are independent of the arrival and service times. Then*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{l}(FCFS) \leq_{E_2} \boldsymbol{l}(\pi) \leq_{E_2} \boldsymbol{l}(LCFS) \tag{3.7}$$

Proof. We have

$$\boldsymbol{l}(\pi) = (W_1(\pi) - r_1, W_2(\pi) - r_2, \cdots, W_N(\pi) - r_N),$$

where $r_1, \cdots, r_N$ are i.i.d random relative due dates being independent of the waiting time variables. Appealing to the closure (under convolution) property of the $\leq_{E_2}$ ordering (cf. [16, Proposition F.6.a, p. 314]) and making use of Proposition 2.1 and (3.6) readily yield (3.7). ∎

Theorem 3.2 *Assume that for any fixed sequence of arrival times $\mathcal{A} = \{a_n\}_{n=1}^N$, the due dates are stochastically comparable in the sense of $\leq_{st}$, viz., for any $m, n \geq 1$, either $d_m \leq_{st} d_n$ or $d_n \leq_{st} d_m$. Then the SSDD (resp. SLDD) discipline minimizes (resp. maximizes) the vector of waiting times within the class Ψ_0 in the sense of $\leq_{E_3}$ ordering:*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{l}(SSDD) \leq_{E_3} \boldsymbol{l}(\pi) \leq_{E_3} \boldsymbol{l}(SLDD) \tag{3.8}$$

Proof. The scheme of the proof is similar to that of Theorem 3.1. We will only consider the first inequality of (3.10). The second one can be shown in an analogous way. Let $f : I\!R \to I\!R$ be an arbitrary convex function and the arrival times $\mathcal{A} = \{a_n\}_{n=1}^N$ and service times $\mathcal{S} = \{\sigma_n\}_{n=1}^N$ be arbitrarily fixed.

Consider an arbitrary discipline $\pi \in \Psi_0$ defined on the sequence of service times $\mathcal{S}$. If π is SSDD, then we are done. Otherwise, let m, $1 \leq m \leq N-1$, be the first time that π does not follow the SSDD rule, i.e., $\pi_i = SSDD_i$, $1 \leq i \leq m-1$. Suppose that at the m-th assignment, π selects customer j whereas customer k would be chosen for service if the SSDD rule were applied: $d_j \geq_{st} d_k$, and $\pi_m = j$, $SSDD_m = k$. Suppose further that $\pi_n = k$, namely, customer k is the n-th that starts service under π, $m < n \leq N$, which implies that $b_j(\pi) \leq b_k(\pi)$.

Construct now a non-idling discipline π' which differs from π only in the customer selections of j and k, viz.,

$$\pi'_m = \pi_n, \qquad \pi'_n = \pi_m, \qquad \pi'_i = \pi_i, \quad i \neq m, \ \ i \neq n, \ \ 1 \leq i \leq N.$$

The discipline π' operates on the sequence of service times $\mathcal{S}' = \{\sigma'_i\}_{i=1}^N$, with

$$\sigma'_j = \sigma_k, \qquad \sigma'_k = \sigma_j, \qquad \sigma'_i = \sigma_i, \quad i \neq j, \ \ i \neq k, \ \ 1 \leq i \leq N.$$

One can verify that $\mathcal{S}'$ is equivalent to $\mathcal{S}$ in law, and independent of the arrival and due dates (cf. [15]), and that

$$b_j(\pi') = b_k(\pi), \qquad b_k(\pi') = b_j(\pi), \qquad b_i(\pi') = b_i(\pi), \quad i \neq j, \;\; i \neq k, \;\; 1 \leq i \leq N,$$

so that π' is feasible.

Define

$$\Lambda_{\mathcal{A},\mathcal{S}}(\pi, f) = E_{\pi,\mathcal{A},\mathcal{S}}\left[\sum_{i=1}^{N} f(l_i(\pi))\right].$$

Observe first that

$$\begin{aligned}
&\Lambda_{\mathcal{A},\mathcal{S}}(\pi, f) - \Lambda_{\mathcal{A},\mathcal{S}'}(\pi', f) \\
&= E_{\pi,\mathcal{A},\mathcal{S}}[f(l_j(\pi))] + E_{\pi,\mathcal{A},\mathcal{S}}[f(l_k(\pi))] - E_{\pi',\mathcal{A},\mathcal{S}'}[f(l_j(\pi'))] - E_{\pi',\mathcal{A},\mathcal{S}'}[f(l_k(\pi'))] \\
&= E_{\pi,\mathcal{A},\mathcal{S}}[f(b_j(\pi) - d_j)] + E_{\pi,\mathcal{A},\mathcal{S}}[f(b_k(\pi) - d_k)] \\
&\quad - E_{\pi',\mathcal{A},\mathcal{S}'}[f(b_k(\pi') - d_k)] - E_{\pi',\mathcal{A},\mathcal{S}'}[f(b_j(\pi') - d_j)] \\
&= E_{\pi,\mathcal{A},\mathcal{S}}[f(b_j(\pi) - d_j)] + E_{\pi,\mathcal{A},\mathcal{S}}[f(b_k(\pi) - d_k)] \\
&\quad - E_{\pi,\mathcal{A},\mathcal{S}'}[f(b_j(\pi) - d_k)] - E_{\pi,\mathcal{A},\mathcal{S}'}[f(b_k(\pi) - d_j)].
\end{aligned}$$

Applying Strassen's theorem to the random variables d_j, d_k entails that there are two random variables $\hat{d}_j, \hat{d}_k$ on a common probability space such that

$$\hat{d}_j =_{st} d_j, \qquad \hat{d}_k =_{st} d_k, \quad \text{and} \quad \hat{d}_j \geq \hat{d}_k, \quad a.s.$$

Due to the assumption that the sequence $\mathcal{D} = \{d_k\}_{k=1}^{N}$ is independent of the sequences of arrival and service times, we get that

$$\begin{aligned}
&\Lambda_{\mathcal{A},\mathcal{S}}(\pi, f) - \Lambda_{\mathcal{A},\mathcal{S}'}(\pi', f) \\
&= E_{\pi,\mathcal{A},\mathcal{S}}[f(b_j(\pi) - \hat{d}_j)] + E_{\pi,\mathcal{A},\mathcal{S}}[f(b_k(\pi) - \hat{d}_k)] \\
&\quad - E_{\pi,\mathcal{A},\mathcal{S}'}[f(b_j(\pi) - \hat{d}_k)] - E_{\pi,\mathcal{A},\mathcal{S}'}[f(b_k(\pi) - \hat{d}_j)].
\end{aligned}$$

The facts that $b_j(\pi) \leq b_k(\pi)$ and that the function f is convex immediately imply

$$f(b_j(\pi) - \hat{d}_j) + f(b_k(\pi) - \hat{d}_k) \geq f(b_j(\pi) - \hat{d}_k) + f(b_k(\pi) - \hat{d}_j).$$

Therefore,

$$\Lambda_{\mathcal{A},\mathcal{S}}(\pi, f) - \Lambda_{\mathcal{A},\mathcal{S}'}(\pi', f) \geq 0.$$

Repeating this interchange process for at most N times yields

$$\Lambda_{\mathcal{A},\mathcal{S}}(\pi, f) \geq \Lambda_{\mathcal{A},\mathcal{S}''}(SSDD, f)$$

Table 1: Comparison Results on the Lag Times and Waiting Times

performance metrics	due date	relative due date	service time distribution	best discipline	worst discipline	stochastic ordering
lag time	known	—	general	SDD	LDD	E_1
lag time	unknown	i.i.d. r.v.	general	FCFS	LCFS	E_2
lag time	$\leq_{st}$ comparable	—	general	SSDD	SLDD	E_3
lag time	unknown	$\leq_{st}$ increasing	general	FCFS	LCFS	E_3
waiting time	—	—	general	FCFS	LCFS	E_1

for some permutation $\mathcal{S}''$ of $\mathcal{S}$. Unconditioning with respect to $\mathcal{A}$ and $\mathcal{S}$ entails that for any convex function $f: \ I\!R \to I\!R$,

$$\sum_{i=1}^{N} E[f(l_i(\pi))] \geq \sum_{i=1}^{N} E[f(l_i(SSDD))],$$

provided the expectations exist. ∎

Corollary 3.2 *Assume that the relative due dates are independent, stochastically increasing random variables that are independent of the arrival and service times. Assume further that the due dates are unknown a priori. Then*

$$\forall \pi \in \Psi_0: \qquad \boldsymbol{l}(FCFS) \leq_{E_3} \boldsymbol{l}(\pi) \leq_{E_3} \boldsymbol{l}(LCFS) \tag{3.9}$$

Proof. Under these assumptions, the SSDD and SLDD disciplines coincide with the FCFS and LCFS. Applying Theorem 3.2 immediately yields the desired result. ∎

The results obtained in this subsection can be summarized in Table 1.

3.2 Latenesses and Response Times

Theorem 3.3 *Assume that for any fixed sequence of arrival times $\mathcal{A} = \{a_n\}_{n=1}^{N}$, the due dates are stochastically comparable in the sense of $\leq_{st}$. Then the SSDD (resp. SLDD) discipline minimizes (resp. maximizes) the vector of latenesses within the class Ψ_0 in the sense of $\leq_{E_3}$ ordering:*

$$\forall \pi \in \Psi_0: \qquad \boldsymbol{L}(SSDD) \leq_{E_3} \boldsymbol{L}(\pi) \leq_{E_3} \boldsymbol{L}(SLDD) \tag{3.10}$$

Proof. Observe that for any customer n, $1 \leq n \leq N$, the lag time $l_n(\pi) = b_n(\pi) - d_n$ is independent of its service time σ_n. Since $L_n(\pi) = l_n(\pi) + \sigma_n$, the relation (3.10) follows from Theorem 3.2 and the closure (under convolution) property of the $\leq_{E_3}$ ordering (cf. [16, Proposition F.6.a, p. 314]). ∎

Remark: In the single server queue, one can show by a coupling argument (see Baccelli et al. [2]) that the E_1 ordering holds:

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{L}(SDD) \leq_{E_1} \boldsymbol{L}(\pi) \leq_{E_1} \boldsymbol{L}(LDD).$$

When the relative due dates are independent and stochastically increasing random variables, the SSDD and SLDD disciplines coincide with FCFS and LCFS, respectively. Applying Theorem 3.3 implies:

Corollary 3.3 *Assume that the due dates are unknown a priori, and that the relative due dates are independent, stochastically increasing random variables that are independent of the arrival and service times. Then*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{L}(FCFS) \leq_{E_3} \boldsymbol{L}(\pi) \leq_{E_3} \boldsymbol{L}(LCFS) \tag{3.11}$$

Setting the due dates to be the arrival times in Corollary 3.3 immediately yields:

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{R}(FCFS) \leq_{E_3} \boldsymbol{R}(\pi) \leq_{E_3} \boldsymbol{R}(LCFS) \tag{3.12}$$

Stronger orderings can be obtained in $G/D/s$ or $G/M/s$ systems where the service times are deterministic or i.i.d. exponentially distributed r.v.'s, respectively.

Theorem 3.4 *Assume the queueing system is $G/D/s$ or $G/M/s$. If the due dates are known, then the SDD (resp. LDD) discipline minimizes (resp. maximizes) the vector of latenesses within the class Ψ_0 in the sense of $\leq_{E_1}$ ordering:*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{L}(SDD) \leq_{E_1} \boldsymbol{L}(\pi) \leq_{E_1} \boldsymbol{L}(LDD). \tag{3.13}$$

Proof. In the case of the $G/D/s$ queueing system, the proof is analogous to that of Theorem 3.1, using in addition the fact that majorization is preserved when all components of the vectors are increased by the same fixed amount. In the case of the $G/M/s$ queueing system, the assertion can be shown with the same argument as that used in the proof of Theorem 4.1 below. The detailed proofs are omitted. ∎

Table 2: Comparison Results on the Latenesses and Response Times

performance metrics	due date	relative due date	service time distribution	best discipline	worst discipline	stochastic ordering
lateness	known	—	Dirac	SDD	LDD	E_1
lateness	known	—	exponential	SDD	LDD	E_1
lateness	unknown	i.i.d. r.v.	Dirac	FCFS	LCFS	E_2
lateness	unknown	i.i.d. r.v.	exponential	FCFS	LCFS	E_2
lateness	$\leq_{st}$ comparable	—	general	SSDD	SLDD	E_3
lateness	unknown	$\leq_{st}$ increasing	general	FCFS	LCFS	E_3
response time	—	—	Dirac	FCFS	LCFS	E_1
response time	—	—	exponential	FCFS	LCFS	E_1
response time	—	—	general	FCFS	LCFS	E_3

Setting the due dates to be the arrival times in the above result implies extremal properties of the FCFS and LCFS disciplines on response times: in the $G/D/s$ or $G/M/s$ queueing systems:

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{R}(FCFS) \leq_{E_1} \boldsymbol{R}(\pi) \leq_{E_1} \boldsymbol{R}(LCFS). \tag{3.14}$$

Using the same argument as the proof of Corollary 3.1 we obtain

Corollary 3.4 *Assume the queueing system is $G/D/s$ or $G/M/s$. If the due dates are unknown a priori, and if the relative due dates are i.i.d. random variables which are independent of the arrival and service times. Then*

$$\forall \pi \in \Psi_0 : \qquad \boldsymbol{L}(FCFS) \leq_{E_2} \boldsymbol{L}(\pi) \leq_{E_2} \boldsymbol{L}(LCFS) \tag{3.15}$$

The results obtained in this subsection can be summarized in Table 2.

3.3 End-to-End Delays

Theorem 3.5 *The FCFS (resp. LCFS) discipline minimizes (resp. maximizes) the end-to-end delays in the sense of stochastic ordering within the class of non-idling non-preemptive disciplines:*

$$\forall \pi \in \Psi_0, \quad \forall n,\ 1 \leq n \leq N : \qquad D_n(FCFS) \leq_{st} D_n(\pi) \leq_{st} D_n(LCFS). \tag{3.16}$$

Proof. We consider the FCFS discipline first. Let π be an arbitrary non-idling non-preemptive discipline. For any fixed n, $1 \leq n \leq N$, we will show that there is a probability space such that

$$q_n(FCFS) \leq q_n(\pi) \quad a.s. \tag{3.17}$$

We arbitrarily fix the arrival times $\{a_i\}_{i=1}^N$. Denote by $\{\sigma_i^\pi\}_{i=1}^N$ the sequence of service times of policy π. Let $\mathcal{I} = \{i_j : 1 \le j \le n\}$ be the set of scheduling indices under π such that $i_1 < i_2 < \ldots < i_n$ and $\pi_{i_j} \in \{1, \ldots, n\}$. Define a new policy γ that schedules customers $n+1, n+2, \cdots, N$ according to π and switches the relative service order of the first n customers to FCFS,

$$\begin{aligned} \gamma_{i_j} &= j, \quad j = 1, 2, \ldots, n, \\ \gamma_l &= \pi_l, \quad l \notin \mathcal{I} \end{aligned}$$

The sequence of service times of policy γ, $\{\sigma_i^\gamma\}_{i=1}^N$, is defined on the same probability space in such a way that the service times are accordingly switched:

$$\sigma_j^\gamma = \begin{cases} \sigma_{\pi_{i_j}}^\pi, & j = 1, \ldots, n, \\ \sigma_j^\pi, & j = n+1, \ldots, N \end{cases}$$

Note that the service times in $\{\sigma_i^\gamma\}_{i=1}^N$ are still i.i.d. and have the same distribution as those in $\{\sigma_i^\pi\}_{i=1}^N$.

It is easy to see that under such a construction, the sequences of increasingly ordered service completion times are the same under π and γ. Moreover, $q_n(\gamma) = q_n(\pi)$.

Now, consider FCFS. We construct again the sequence of service times of the FCFS policy on the same probability space and define it to be identical to that of γ: $\sigma_i^{FCFS} = \sigma_i^\gamma$, $1 \le i \le N$. It should be clear that $c_i(FCFS) \le c_i(\gamma)$, $1 \le i \le n$. (Observe that the service completion times under FCFS may or may not coincide with those under γ.) Therefore, $q_n(FCFS) \le q_n(\gamma)$, so that relation (3.17) holds. Applying Strassen's theorem (cf. Lemma 2.1) immediately yields the first inequality of (3.16).

Consider now the LCFS part. The idea of the proof is similar. For any fixed n, $1 \le n \le N$, we will show that there is a probability space such that

$$q_n(LCFS) \ge q_n(\pi) \quad a.s. \tag{3.18}$$

As in the previous case, we arbitrarily fix the arrival times $\{a_i\}_{i=1}^N$, and the service times $\{\sigma_i^\pi\}_{i=1}^N$ of policy π. Let ρ be a new policy whose service times are defined on the same probability space and identical to those of π: $\{\sigma_i^\rho\}_{i=1}^N = \{\sigma_i^\pi\}_{i=1}^N$. Define ρ to be such that the LCFS rule is applied to the customers $n+1, n+2, \cdots, N$. Due to the fact that in ρ, customers $n+1, n+2, \cdots, N$ have higher priority than those of $1, \cdots, n$, we have the inequalities: $c_i(\rho) \ge c_i(\pi)$, $1 \le i \le n$. Therefore, $q_n(\rho) \ge q_n(\pi)$.

Consider now policy ρ. We examine the scheduling of customers $1, \cdots, n$. We interchange, whenever necessary, their scheduling positions as well as their service times according to the LCFS rule. Since the sequence of the increasingly ordered service completion times remains the same (due to

the interchange of service times), the resulting policy is LCFS. Further, $q_n(LCFS) = q_n(\rho)$. Hence, relation (3.18) holds. An application of Strassen's theorem readily implies the second inequality of (3.16). ∎

4 Comparisons within the Class of Non-Idling and Preemptive Service Disciplines

In this section, we generalize the results obtained in the previous section to the class of non-idling preemptive disciplines. Throughout this section, we assume that the service times are i.i.d. random variables with an exponential distribution. Such an assumption allows us to place some restrictions on the class of service disciplines that we need to consider.

Lemma 4.1 *For any function of the state variables $\boldsymbol{R}$, $\boldsymbol{L}$ and $\boldsymbol{T}$, there exists an optimal discipline, that minimizes or maximizes the expectation of this function, whose decision points occur only at the arrival times and all but the last service completion time.*

In other words, preemptions and new customer assignments occur only at the customer arrival instants and at the instants of all but the last service completion. This fact results from the memoryless property of the exponential distribution. Indeed, between these instants the state represented by the existing customers and their remaining service times does not change. The formal proof is left to the interested reader. Throughout this section, we will confine ourselves to the service disciplines having the property given in Lemma 4.1.

The results obtained in this section are summarized in Table 3. Result 1 of Table 3 are restated and proved in Theorem 4.1 below. Result 2 is a trivial corollary of Theorem 4.1. Result 3 follows from Result 2 and the convolution theorem for the $\leq_{E_2}$ ordering (cf. [16, Proposition F.6.a, p. 314]). Result 4 can be established by using the ideas contained in the proofs of Theorems 3.3 and 4.1. Result 5 is a consequence of Result 4. Finally, a proof of Result 6 can be obtained by making use of the ideas in Theorems 3.5 and 4.1.

Theorem 4.1 *Assume that the due dates are known. Then the PSDD (resp. PLDD) discipline minimizes (resp. maximizes) the vector of latenesses within the class non-idling and preemptive disciplines Ψ_{ni} in the sense of $\leq_{E_1}$ ordering:*

$$\forall \pi \in \Psi_{ni}: \qquad \boldsymbol{L}(PSDD) \leq_{E_1} \boldsymbol{L}(\pi) \leq_{E_1} \boldsymbol{L}(PLDD). \tag{4.1}$$

Proof. We consider the system as if each server is continually serving customers. Whenever a service completion occurs and there is no customer assigned to that server, it corresponds to the

Table 3: Comparison Results within Preemptive Disciplines

	performance metrics	due date	relative due date	best discipline	worst discipline	stochastic ordering
1	lateness	known	—	PSDD	PLDD	E_1
2	response time	—	—	FCFS	PLCFS	E_1
3	lateness	unknown	i.i.d. r.v.	FCFS	PLCFS	E_2
4	lateness	$\leq_{st}$ comparable	—	PSSDD	PSLDD	E_3
5	lateness	unknown	$\leq_{st}$ increasing	FCFS	PLCFS	E_3
6	end-to-end delay	—	—	FCFS	PLCFS	$\leq_{st}$

completion of a fictitious customer. When a customer is assigned to a server, it is assigned a service time equal to the remainder of the service time already underway at that server. The exponential assumption guarantees that the customer service times are i.i.d. exponential r.v.'s.

Consider the result pertaining to PSDD (a similar argument can be used to obtain the desired result for PLDD). Assume that the arrival times and service times are given. Let $0 = t_1 < t_2 < \cdots < t_{2N}$ be the decision epochs (note that there are at most $2N$ decision epochs). Let π be an arbitrary service discipline that is not PSDD. We will construct a new discipline π' that violates the PSDD rule one less time and decreases the lateness vector in the sense of E_1 ordering.

Let m, $1 \leq m \leq 2N-1$, be the first time that π does not follow the PSDD rule, i.e., at time t_m, there exist customers j and k in the system with $d_j > d_k$ and that customer j is assigned to a server at time t_m but not customer k.

The policy π' is constructed as follows. The (residual) service times under π' are, as under π, associated with the servers, and are the same as those with π. The decisions of π' are defined as follows. For all $1 \leq n < m$, the decisions of π' at time t_n is the same as π. At time t_m, π' assigns customer k in place of j to a server and maintains the same assignment for all other customers. If at time t_{m+1}, customer j finishes under π, which implies that customer k finishes under π', then for all $m < n \leq 2N$, the assignment decisions of π at time t_n are the same as π except that when customer k is assigned to a server under π, the customer j will be assigned to the server under π'. Otherwise, if customer j does not finish under π at time t_{m+1} (nor does customer k under π'), then for all $m < n \leq 2N$, the assignment decisions of π at time t_n are exactly the same as π (even for customers j and k).

In both cases, one easily verify (using the inequality $d_j > d_k$) that $\boldsymbol{L}(\pi') \prec \boldsymbol{L}(\pi)$. This can be performed repeatedly to produce a service discipline where the PSDD rule is applied everywhere so that $\boldsymbol{L}(PSDD) \prec \boldsymbol{L}(\pi)$. Removal of the conditioning on the arrival times and service times yields the desired result for PSDD. ∎

5 Optimality of Non-Idling FCFS Service Disciplines

In this section, we establish the optimality of non-idling FCFS service disciplines compared with idling disciplines. The following lemma provides the basis of our proofs in this section and has independent interest.

Lemma 5.1 *For any non-preemptive discipline $\pi \in \Psi_{np}$, and any non-idling non-preemptive discipline $\pi' \in \Psi_0$, we have*

$$\left(b_{\pi'_1}(\pi'), b_{\pi'_2}(\pi'), \cdots, b_{\pi'_N}(\pi')\right) \leq_{st} \left(b_{\pi_1}(\pi), b_{\pi_2}(\pi), \cdots, b_{\pi_N}(\pi)\right). \tag{5.1}$$

$$\left(c_{\pi'_1}(\pi'), c_{\pi'_2}(\pi'), \cdots, c_{\pi'_N}(\pi')\right) \leq_{st} \left(c_{\pi_1}(\pi), c_{\pi_2}(\pi), \cdots, c_{\pi_N}(\pi)\right). \tag{5.2}$$

Proof. Owing to the fact that the service times are i.i.d. random variables, we can couple the service times in such a way that the service time of the n-th assigned customer in both π and π' is σ_n.

Since π is non-preemptive, we have the following recursive equation concerning the customer service beginning and completion times:

$$b_{\pi_n}(\pi) = \max\left(a_{\pi_n}, c_{\gamma_{n-s(\pi,n)}}\right) + \xi_{\pi_n}(\pi), \tag{5.3}$$

$$c_{\pi_n}(\pi) = b_{\pi_n}(\pi) + \sigma_n, \tag{5.4}$$

where, $\gamma_i(\pi, n)$ is the index of the i-th largest number in $\{c_{\pi_1}(\pi), c_{\pi_2}(\pi), \cdots, c_{\pi_{n-1}}(\pi)\}$, $\xi_n(\pi) \geq 0$ denotes the time interval between the epoch that one of the servers is available for the service of customer π_n after its arrival and the epoch when the customer starts service under (idling) discipline π. By convention, $c_i = 0$ and $\gamma_i(\pi, n) = 0$ for all $i \leq 0$.

For the non-idling and non-preemptive discipline π', we have that

$$b_{\pi'_n}(\pi') = \max\left(a_{\pi'_n}, c_{\gamma_{n-s(\pi',n)}}\right), \tag{5.5}$$

$$c_{\pi'_n}(\pi') = b_{\pi'_n}(\pi') + \sigma_n. \tag{5.6}$$

It is thus clear from equations (5.3—5.6) that

$$\forall n, \ 1 \leq n \leq N: \qquad b_{\pi'_n}(\pi') \leq b_{\pi_n}(\pi),$$

and that

$$\forall n, \ 1 \leq n \leq N: \qquad c_{\pi'_n}(\pi') \leq c_{\pi_n}(\pi),$$

which imply the relations (5.1) and (5.2), respectively. ∎

Corollary 5.1 *For any non-preemptive discipline $\alpha \in \Pi_{FCFS}$, we have*

$$(b_1(FCFS), b_2(FCFS), \cdots, b_N(FCFS)) \leq_{st} (b_1(\alpha), b_2(\alpha), \cdots, b_N(\alpha)), \tag{5.7}$$

$$(c_1(FCFS), c_2(FCFS), \cdots, c_N(FCFS)) \leq_{st} (c_1(\alpha), c_2(\alpha), \cdots, c_N(\alpha)). \tag{5.8}$$

Proof. It suffices to note that for all $\alpha \in \Pi_{FCFS}$, $\alpha_n = n$. ∎

We are now in a position to prove the optimality of the FCFS discipline within the class of non-preemptive idling disciplines. We consider the customer end-to-end delays in Theorem 5.1 below. Other optimality properties of the FCFS are presented in Table 4 (Results 1—8) and can be shown in an analogous way.

Theorem 5.1 *FCFS minimizes the end-to-end delays in the sense of stochastic ordering, within the class of non-preemptive idling disciplines:*

$$\forall \pi \in \Psi_{np}, \quad \forall n,\ 1 \leq n \leq N: \qquad D_n(FCFS) \leq_{st} D_n(\pi). \tag{5.9}$$

Proof. Observe first that for all $\pi \in \Psi_{np}$, there exists $\alpha \in \Pi_{FCFS}$ such that

$$\forall n,\ 1 \leq n \leq N: \qquad q_n(\alpha) \leq_{st} q_n(\pi). \tag{5.10}$$

This last relation can be shown by mimicking the proof of Theorem 3.5. It consists in constructing the new discipline α according to π in such a way that α has the same decision times and the same idling times as π.

The assertion of the theorem becomes now a consequence of Proposition 2.1 and Corollary 5.1 together with relation (5.10). ∎

Remark: Result 5 in Table 4 can also be obtained from the Schur convex ordering established by Foss [9, 10].

Consider now the optimality properties of FCFS within the class of preemptive idling disciplines Ψ: Results 9—12 in Table 4. The proof of these results rely on the following lemma.

Lemma 5.2 *Assume that the service times are i.i.d. r.v.'s with exponential distribution. For all discipline $\pi \in \Psi$, there is a non-idling discipline $\rho \in \Psi_{ni}$ such that*

$$(c_1(\rho), \cdots, c_N(\rho)) \leq_{st} (c_1(\pi), \cdots, c_N(\pi)), \tag{5.11}$$

$$\boldsymbol{L}(\rho) \leq_{st} \boldsymbol{L}(\pi), \tag{5.12}$$

$$\boldsymbol{R}(\rho) \leq_{st} \boldsymbol{R}(\pi). \tag{5.13}$$

Table 4: Optimality of FCFS within Idling Disciplines

	class of disciplines	performance metrics	due date	relative due date	service time distribution	stochastic ordering
1	non-preemptive	waiting time	—	—	general	$E_1^{\uparrow}$
2	non-preemptive	lag time	unknown	i.i.d. r.v.	general	$E_2^{\uparrow}$
3	non-preemptive	lag time	unknown	$\leq_{st}$ increasing	general	$E_3^{\uparrow}$
4	non-preemptive	response time	—	—	Dirac	$E_1^{\uparrow}$
5	non-preemptive	response time	—	—	general	$E_3^{\uparrow}$
6	non-preemptive	lateness	unknown	i.i.d. r.v.	Dirac	$E_2^{\uparrow}$
7	non-preemptive	lateness	unknown	$\leq_{st}$ increasing	general	$E_3^{\uparrow}$
8	non-preemptive	end-to-end delay	—	—	general	$\leq_{st}$
9	preemptive	response time	—	—	exponential	$E_1^{\uparrow}$
10	preemptive	lateness	unknown	i.i.d. r.v.	exponential	$E_2^{\uparrow}$
11	preemptive	lateness	unknown	$\leq_{st}$ increasing	exponential	$E_3^{\uparrow}$
12	preemptive	end-to-end delay	—	—	exponential	$\leq_{st}$

Proof. We consider the system as if each server was continually serving customers. Whenever a (virtual) service completion occurs and there is no customer assigned to that server, it corresponds to the completion of a fictitious customer. When a customer is assigned to a server, it is assigned a service time equal to the remainder of the service time already underway at that server. The exponential assumption guarantees that the customer service times are i.i.d. exponential r.v.'s. Assume that the arrival times and service times are given. We show that in such a probability space, there is a discipline ρ which is non-idling and that

$$(c_1(\rho), c_2(\rho), \cdots, c_N(\rho) \leq (c_1(\pi), c_2(\pi), \cdots, c_N(\pi) \quad a.s., \tag{5.14}$$

which will imply the assertions of the lemma.

Let $0 \leq s_1 < s_2 < \cdots$ be the superposition of the sequences of arrival times and of the (virtual) service completion times in the system. Define discipline ρ as follows: For $m = 1, 2, \cdots$,

- If a customer n, $1 \leq n \leq N$, is assigned to a server at s_m under π, and if this customer is not finished by s_m under ρ, then n is also assigned to a server under ρ.
- If there is a server such that no customer is assigned to at time s_m under π, or if the customer that is assigned to at s_m under π is already finished by time s_m under ρ, then ρ assigns a customer waiting in the queue, if any, to that server.

Under this construction, it is easy to see that ρ is non-idling and that relation (5.14) holds. ∎

Results 9—12 in Table 4 now follow from Lemma 5.2 and Proposition 2.1, together with Results 2, 3, 5 and 6 in Table 3.

Remark: When preemptions are allowed, Hirayama and Kijima [11] obtained that in the $G/IFR/1$ model (i.e., the service times are of IFR type),

$$\forall \pi \in \Psi : \qquad \boldsymbol{R}(FCFS) \leq_{E_3^{\uparrow}} \boldsymbol{R}(\pi).$$

A slightly stronger relation was obtained in Chang and Yao [6]. However, this result does not hold for arbitrary $G/IFR/s$ model, $s \geq 2$. Here is a counterexample. Consider a $G/D/3$ queue with 10 arriving customers. The service times are all 5 and the arrival times are $0, 1, 2, 3, 4, 5, 6, 14, 14.001, 14.002$. Under FCFS, the average response time is 6.0998. The following preemptive schedule gives 6.0 Server 1 serves customer 1 at $t = 0$, customer 2 at $t = 5$, customer 5 at $t = 7$, and customer 8 at $t = 14$. Server 2 serves customer 2 at $t = 1$, customer 4 at $t = 4$, customer 7 at $t = 9$, and customer 9 at $t = 14.001$. Server 3 serves customer 3 at $t = 2$, customer 6 at $t = 7$, and customer 10 at $t = 14.002$. Note that customer 2 was preempted at $t = 4$.

6 Extensions

6.1 Comparisons in the Stationary Regime

The extremal properties of the FCFS, LCFS, SDD, LDD, SSDD, SLDD disciplines can be extended to the stationary regime, provided it exists. To this end, we let N go to ∞ and denote by $W(\pi)$, $R(\pi)$, $l(\pi)$, $L(\pi)$, $D(\pi)$ the limit r.v.'s of the weakly convergent r.v.'s $W_n(\pi)$, $R_n(\pi)$, $l_n(\pi)$, $L_n(\pi)$, $D_n(\pi)$, respectively, when n goes to ∞, provided such convergence exists under the service discipline $\pi \in \Psi$.

Since the stochastic ordering $\leq_{st}$ is preserved for the limit r.v.'s of weakly convergent sequences, the stochastic ordering $\leq_{st}$ for the r.v.'s $D_n(\pi)$ is preserved whenever $D_n(\pi)$ weakly converges to $D(\pi)$ as n goes to ∞.

For other performance metrics, we recall that the sequence of random variables $X_n \in I\!R$, $n \geq 1$, will be said to converge weakly to the random variable X for the class of Borel mappings $\mathcal{F} : I\!R \to I\!R$ in the Cesaro sense (cf. Feller [8, p. 249]) if

$$\forall f \in \mathcal{F} : \qquad \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} E[f(X_n)] = E[f(X)]. \tag{6.1}$$

Observe that for any given family $\mathcal{F}$, a sufficient condition for (6.1) to hold is that X_n couples in finite time with a stationary and ergodic sequence. Such a coupling exists, for example, for the waiting time and response times of the non-idling FCFS discipline when the arrival process is stationary (cf. [5, Theorem 5.5.7 and Lemma 5.5.8]).

Lemma 6.1 *Let* $\{X_n\}_{n=1}^{\infty}$ *and* $\{Y_n\}_{n=1}^{\infty}$ *be two sequences of r.v.'s such that*

$$\forall n \geq 1 : \qquad (X_1, \cdots, X_n) \leq_{E_3} \quad (resp. \leq_{E_3^{\uparrow}}) \quad (Y_1, \cdots, Y_n).$$

If the sequence $\{X_n\}$ *and* $\{Y_n\}$ *converge weakly to* X *and* Y *with respect to the class of convex functions (resp. increasing convex functions) in the Cesaro sense when* n *tends to* ∞*, then*

$$X \leq_{cx} Y \qquad (resp.\ X \leq_{icx} Y).$$

Proof. For all convex functions $f : I\!R \to I\!R$, the weak convergence assumptions yield

$$E[f(X))] = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[f(X_i)] \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E[f(Y_i)] = E[f(Y)].$$

■

This last lemma entails that the stochastic ordering $\leq_{E_3}$ and $\leq_{E_3^{\uparrow}}$ (recall that $\leq_{E_1} \Rightarrow \leq_{E_2} \Rightarrow \leq_{E_3}$ and that $\leq_{E_1^{\uparrow}} \Rightarrow \leq_{E_2^{\uparrow}} \Rightarrow \leq_{E_3^{\uparrow}}$) established in the paper on the transient performance metrics $\boldsymbol{W}(\pi)$, $\boldsymbol{R}(\pi)$, $\boldsymbol{l}(\pi)$ and $\boldsymbol{L}(\pi)$ reduce to the stochastic orderings $\leq_{cx}$ and $\leq_{icx}$ on the corresponding stationary performance metrics $W(\pi)$, $R(\pi)$, $l(\pi)$ and $L(\pi)$, respectively, provided the weak convergence assumptions with respect to the convex functions, and the increasing and convex functions, are fulfilled.

Table 5 summarizes the comparison results in steady state where "G" stands for "general" and "E" for "exponential". In Results 4—7 and 10—12 of Table 5, "E" corresponds to Ψ_{ni} and Ψ, respectively, i.e., the service times should be exponentially distributed when preemption is allowed.

6.2 Other Performance Metrics

Denote by $\theta(\pi) = P[L(\pi) \leq 0]$ the *goodput* of discipline $\pi \in \Psi$, i.e., the proportion of customers that finish service by their deadlines. Result 6 of Table 5 implies

Corollary 6.1 *Assume that the due dates are unknown a priori and that the relative due dates are i.i.d. r.v.'s with concave distribution. Then, in a* $G/GI/s$ *queueing system,*

$$\forall \pi \in \Psi_0, \qquad \theta(LCFS) \geq \theta(\pi) \geq \theta(FCFS),$$

and in a $G/M/s$ *queueing system,*

$$\forall \pi \in \Psi_{ni}, \qquad \theta(PLCFS) \geq \theta(\pi) \geq \theta(FCFS).$$

Table 5: Comparison Results in Steady State

	class	performance metrics	due date	relative due date	service time	best discipline	worst discipline	stochastic ordering
1	Ψ_0	lag time	$\leq_{st}$ comparable	—	G	SSDD	SLDD	$\leq_{cx}$
2	Ψ_0	lag time	unknown	$\leq_{st}$ increasing	G	FCFS	LCFS	$\leq_{cx}$
3	Ψ_0	waiting time	—	—	G	FCFS	LCFS	$\leq_{cx}$
4	Ψ_0, Ψ_{ni}	lateness	$\leq_{st}$ comparable	—	G, E	(P)SSDD	(P)SLDD	$\leq_{cx}$
5	Ψ_0, Ψ_{ni}	lateness	unknown	$\leq_{st}$ increasing	G, E	FCFS	(P)LCFS	$\leq_{cx}$
6	Ψ_0, Ψ_{ni}	response time	—	—	G, E	FCFS	(P)LCFS	$\leq_{cx}$
7	Ψ_0, Ψ_{ni}	end-to-end delay	—	—	G, E	FCFS	(P)LCFS	$\leq_{st}$
8	Ψ_{np}	lag time	unknown	$\leq_{st}$ increasing	G	FCFS	—	$\leq_{icx}$
9	Ψ_{np}	waiting time	—	—	G	FCFS	—	$\leq_{icx}$
10	Ψ_{np}, Ψ	lateness	unknown	$\leq_{st}$ increasing	G, E	FCFS	—	$\leq_{icx}$
11	Ψ_{np}, Ψ	response time	—	—	G, E	FCFS	—	$\leq_{icx}$
12	Ψ_{np}, Ψ	end-to-end delay	—	—	G, E	FCFS	—	$\leq_{st}$

Proof. Note that $\theta(\pi) = P[L(\pi) \leq 0] = P[R(\pi) \leq r]$, where r is the relative due date. Thus, if r has a concave distribution, the goodput $\theta(\pi)$ is a convex function of the response time. Therefore, applying Result 6 of Table 5 yields the assertion of the corollary. ∎

In the literature, the *tardinesses* of customers are also analyzed. Let $T_n = \max(0, L_n)$ be the tardiness of customer n, and $\boldsymbol{T} = (T_1, \cdots, T_N)$. Observe that for $x \in I\!R$, the function $f(x) = \max(0, x)$ is increasing and convex, and that the composition of an increasing and (Schur) convex function with such a function f is still increasing and (Schur) convex. Therefore, the stochastic orderings $\leq_{E_i}$ and $\leq_{E_i^{\uparrow}}$ obtained above on the vectors of latenesses $\boldsymbol{L}$ imply the stochastic orderings $\leq_{E_i^{\uparrow}}$ on the vectors of tardinesses $\boldsymbol{T}$, $i = 1, 2, 3$.

6.3 Comparisons in Queueing Networks

The extremal properties obtained above do not hold in general queueing networks with multi-server queues. However, some of the results obtained earlier in this paper can be extended to tandem queueing networks. In addition, the extremal disciplines can be applied to the *source* nodes and *sink* nodes, if any, of some queueing networks.

6.3.1 Tandem queueing networks

Consider a tandem network $\mathcal{K}$ consisting of $K \geq 1$ multiple server nodes. N customers arrive to node 1 according to an arbitrary arrival process. Customers departing from node k enter node

$k+1$, $k=1,\ldots,K-1$. Finally, customers depart the system from node K. Let the arrival times to node 1 be given by $\{a_n\}_{k=1}^N$ and the service times at node k, $\{\sigma_{k,n}\}_{n=1}^N$, $1 \leq k \leq K$. Let $\{d_n\}_{n=1}^N$ be the due dates associated with the N customers.

Let Ψ, Ψ_{np}, etc. denote the same classes of policies as before, except over K nodes rather than a single node. Let the SDD and LDD disciplines be the policies that always serve the customer with the smallest and largest due dates, respectively, in all the nodes. Let the FCFS and LCFS disciplines be defined with respect to the arrival times $\{a_n\}_{k=1}^N$ and be applied to all the nodes. Let $L_n(\pi)$ and $D_n(\pi)$ be the response time and end-to-end delay of the n-th customer in the entire network under policy π.

We have the following extension to Theorems 3.4 and 3.5 to tandem queueing networks.

Theorem 6.1 *Assume that in the tandem queueing network, the nodes are either $\cdot/D/s$ or $\cdot/M/s$ and that service times are independent of each other and the arrival times. Then the FCFS (resp. LCFS) discipline minimizes (resp. maximizes) the end-to-end delays in the sense of stochastic ordering within the class Ψ_0:*

$$\forall \pi \in \Psi_0, \quad \forall n, \leq n \leq N: \quad D_n(FCFS) \leq_{st} D_n(\pi) \leq_{st} D_n(LCFS).$$

If the due dates are known, then the SDD (resp. LDD) discipline minimizes (resp. maximizes) the vector of latenesses in the sense of the $\leq_{E_1}$ ordering:

$$\forall \pi \in \Psi_0: \quad \boldsymbol{L}(SDD) \leq_{E_1} \boldsymbol{L}(\pi) \leq_{E_1} \boldsymbol{L}(LDD).$$

The comparisons in this theorem can be established using interchange arguments similar to those used in earlier proofs and the ordering on permutations (see [2]). Similar results can be established in some cases where due dates are not known and for the class of policies Ψ_{ni}. Last, these results can be extended to the stationary regime in which case the underlying orderings become $\leq_{st}$ and $\leq_{cx}$ among the stationary end-to-end delay and lateness, respectively.

6.3.2 Source and sink nodes in general queueing networks

Consider a network $\mathcal{K}$ with $K \geq 1$ nodes. A node consists of a waiting queue and several servers. When a customer finishes service at a node, it is (randomly) routed to one of the successor nodes. The nodes with no predecessors are called the sources, and those having no successors are called the sinks. Let $\mathcal{K}_0$ and $\mathcal{K}_1$ denote the sets of source nodes and sink nodes, respectively. Note that the set $\mathcal{K}_0 \cap \mathcal{K}_1$ may not be empty.

It is assumed that all the customers arrive in the system by one of the source nodes, and that a customer can leave the system only when it finishes service at one of the sink nodes. There is a resequencing buffer with infinite capacity in the system. When a customer leaves a sink node, it

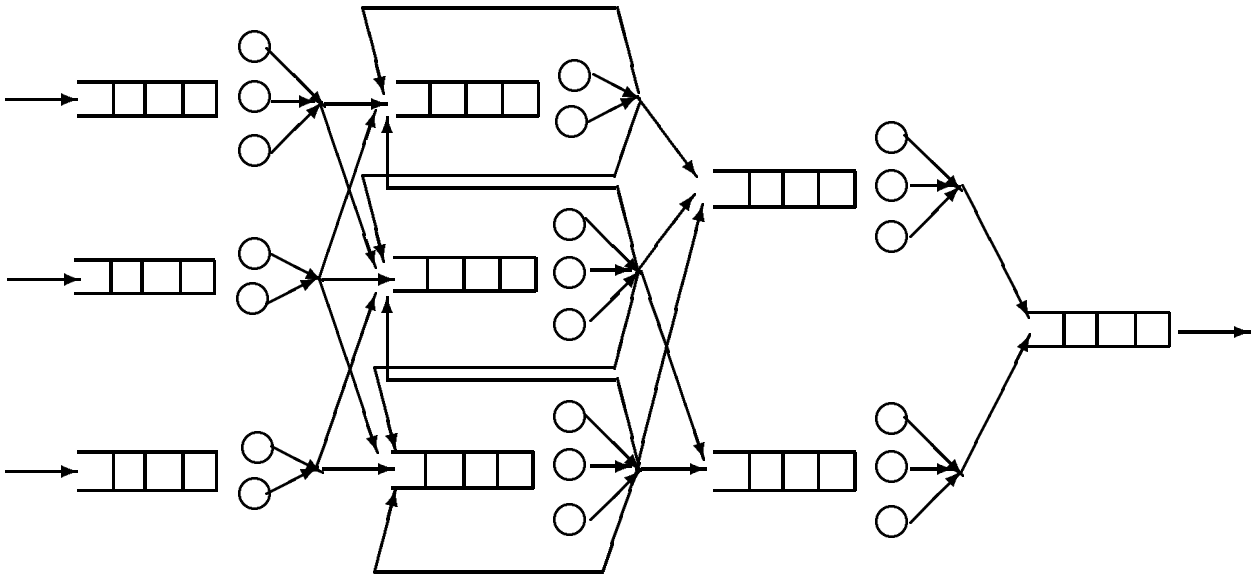

Figure 1: Example of a queueing network containing source and sink queues

enters the resequencing buffer. The resequencing is performed with respect to the global arrival times of the customers to the system. Figure 1 illustrates an example of such queueing networks. A particular case is the queueing system with $K \geq 1$ independent and parallel multi-server queues: $\mathcal{K}_0 = \mathcal{K}_1$.

Let $a_1 = 0 \leq a_2 \leq \cdots \leq a_N$ be the time epochs when customers arrive in the system (i.e., at one of the source queues). For any discipline π, let $b_n(\pi)$ (resp. $c_n(\pi)$) be the service beginning (resp. completion) time of customer n at one of the source (resp. sink) queues. Then, the performance metrics are defined as $l_n(\pi) = b_n(\pi) - d_n$, $W_n(\pi) = b_n(\pi) - a_n$, $L_n(\pi) = c_n(\pi) - d_n$, $R_n(\pi) = c_n(\pi) - a_n$, and $q_n(\pi) = \max_{1 \leq m \leq n} c_m(\pi)$, $D_n(\pi) = q_n(\pi) - a_n$.

If the servers of each source queue are identical, then all the previously established results pertaining to the customer lag times and waiting times hold in the queueing network $\mathcal{K}$, provided the extremal disciplines are applied to all the source queues. Similarly, if the servers of each sink queue are identical, then all the previously established results pertaining to the customer latenesses and response times as well as end-to-end delays hold in the queueing network $\mathcal{K}$, provided the extremal disciplines are applied to all the sink queues.

References

[1] F. Baccelli, E. Gelenbe, B. Plateau, "An End-to-End Approach to the Resequencing Problem", *Journal of the ACM*, Vol. 31, pp. 474-485, 1984.

[2] F. Baccelli, Z. Liu, D. Towsley, "Extremal Scheduling of Parallel Processing with and without Real-Time Constraints", to appear in the *Journal of the ACM*.

[3] F. Baccelli, A. M. Makowski, "Queueing Models for Systems with Synchronization Constraints", *Proceedings of the IEEE*, Vol. 77, Special Issue on Dynamics of Discrete Event Systems, pp. 138-161, Jan. 1989.

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

[5] A. Brandt, P. Franken, B. Lisek, *Stationary Stochastic Models.* Akademic-Verlag, 1989.

[6] C. S. Chang, D. D. Yao, "Rearrangement, Majorization and Stochastic Scheduling", IBM Research Report RC 16250, 1990.

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

[8] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. II, Second Edition, John Wiley & Sons, 1971.

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

[10] 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.*, Vol. 22, pp. 142-147, 1981.)

[11] T. Hirayama, M. Kijima, "An Extremal Property of FIFO discipline in $G/IFR/1$ Queues", *Adv. Appl. Prob.*, Vol. 21, 481-484, 1989.

[12] I. Iliadis, L. Lien, "Resequencing Delay for a Queueing System with Two Heterogeneous Servers Under a Threshold-Type Scheduling", *IEEE Trans. on Communications*, Vol. 36, pp. 692-702, 1988.

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

[14] L. Kleinrock, F. Kamoun, R. Muntz, "Queueing Analysis of the reordering issue in a distributed database concurrency control mechanism", *Proc. of the 2nd International Conference on Distributed Computing Systems*, Versailles, France, 1981.

[15] Z. Liu, P. Nain, "Optimal Scheduling in Some Multi-Queue Single-Server Systems." *IEEE Transactions on Automatic Control*, Vol. 37, pp. 247–252, 1992.

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

[17] R. Nelson, D. Towsley, A. N. Tantawi, "Performance Analysis of Parallel Processing Systems", *IEEE Trans. on Software Engineering*, Vol. 14, No. 4, pp. 532-540, April 1988.

[18] M. Pinedo, "Stochastic Scheduling with Release Dates and Due Dates." *Oper. Res.*, Vol. 31, No. 3, pp. 559-572, May-June 1983.

[19] J. G. Shantikumar, U. Sumita, "Convex Ordering of Sojourn Times in Single-Server Queues: Extremal Properties of FIFO and LIFO Service Disciplines." *J. Appl. Prob.*, Vol. 24, pp. 737-748, 1987.

[20] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models.* English translation (D.J. Daley editor), J.Wiley and Sons, New York, 1984.

[21] V. Strassen, "The existence of Probability Measures with Given Marginals," *Ann. Math. Stat.*, Vol. 36, pp. 423-439, 1965.

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

[23] W. Whitt, "The Amount of Overtaking in a Network of Queues," *Networks*, Vol. 14, pp. 411-426, 1984.

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

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