SRC TR 89-37

Simple Proofs of Some Folk Theorems for Parallel Queues

By

A. M. Makowski
&
T. K. Philips

April 1989
Submitted to QUEUEING SYSTEMS

SIMPLE PROOFS OF SOME FOLK THEOREMS FOR PARALLEL QUEUES

by

Armand M. Makowski[1] and Thomas K. Philips[2]

ABSTRACT

We present simple proofs of some folk theorems for systems of identical single server queues operating in parallel. In particular we establish a monotonicity property in the number of servers, and show that round-robin customer assignment outperforms random customer assignment. The results are couched in terms of stochastic orderings and hold in great generality.

[1] Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742. The work was performed while this author was visiting the IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

[2] IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598.

I. INTRODUCTION

Queueing theory has a large collection of so-called folk theorems, i.e., statements which are intuitively reasonable and whose essential truth is widely accepted by workers in the field, but whose precise range of validity is never fully explored. In fact, the vast majority of existing folk theorems do not seem to have rigorous proofs, and are often stated in statistical equilibrium under restricted model assumptions, such as the exponentiality of inter-arrival and service times.

A typical example of such a folk theorem emerges from the belief that "determinism minimizes waiting time" [4,5]. Although this statement is open to interpretation, it is usually understood as that the mean waiting time in a single server, infinite capacity work-conserving queue is minimized for any fixed utilization by a deterministic interarrival process. Over the years, various formal versions of this folk theorem have appeared in the literature [4,5], with some of them taking the form of stochastic dominance results [1,2,10,11].

In this paper, we examine some folk theorems for systems of identical single server queues operating in parallel. In particular we establish a monotonicity property in the number of servers under the random customer assignment, and show that the round-robin customer assignment outperforms the random customer assignment. The results are couched in terms of stochastic orderings and hold in great generality under minimal assumptions on the distribution of the inter-arrival and service times. The proofs are fairly simple and hold both during the transient phase and in the statistical equilibrium; the arguments are crucially dependent on the recursive nature of the equations satisfied by the quantities of interest, and make use of various properties of stochastic orderings.

The paper is organized as follows: Some basic facts on stochastic orderings are summarized in Section 2, where a simple comparison result is also recalled. This result formalizes the classical statement that "determinism minimizes waiting time" and provides a unified tool for estalishing the aforementioned properties of parallel servers. In Section 3, two different representations are introduced for systems of identical servers operating in parallel under independent customer assignment proceduress. These two representations are then used in Section 3 in order to establish precise statements of the folk theorems.

II. STOCHASTIC ORDERINGS

In this section, we introduce several notions of stochastic orderings and briefly explain their use in generating stochastic comparisons on quantities of interest. We first present the notation and conventions used throughout the paper.

II.1. Notation, conventions and definitions

The set of real (resp. non-negative real) numbers is denoted by $\mathbb{R}$ (resp. $\mathbb{R}_+$). All the random variables (RV's) of interest are defined on some common probability triple $(\Omega, \mathcal{F}, P)$. The k^{th} component RV of any $\mathbb{R}^K$-valued RV X is denoted by X^k, $1 \leq k \leq K$; a similar convention is adopted for the components of any vector in $\mathbb{R}^K$.

We denote by $\mathcal{D}(\mathbb{R}^K)$ the collection of all probability distribution functions on $\mathbb{R}^K$. We identify an element F of $\mathcal{D}(\mathbb{R}^K)$ with an $\mathbb{R}^K$-valued RV $X = (X^1, \ldots, X^K)$ which has distribution F, in which case

$$F(x) = P[X^1 \leq x^1, \ldots, X^K \leq x^K] \tag{2.1}$$

for all $x = (x^1, \ldots, x^K)$ in $\mathbb{R}^K$, and we set

$$m(F) = \int_{\mathbb{R}^n} x dF(x) = m(X) \tag{2.2}$$

whenever the first moment of F exists. Two $\mathbb{R}^K$-valued RV's X and Y are said to be *equal in law* if they have same distribution, a fact we denote by $X =_{st} Y$, in agreement with the notation introduced below.

For any two vectors x and y in $\mathbb{R}^K$, the ordering $x \leq y$ is interpreted componentwise to read $x^k \leq y^k$, $1 \leq k \leq K$. A mapping $f : \mathbb{R}^K \to \mathbb{R}$ is said to be monotone non-decreasing (resp. non-increasing) if $x \leq y$ in $\mathbb{R}^K$ implies $f(x) \leq$ (resp. $\geq$)$f(y)$, while it is convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all λ in the interval $[0,1]$ and every pair of vectors x and y in $\mathbb{R}^K$. Moreover, as customary in the literature on Queueing Theory, we define the mapping $[\,\cdot\,]^+ : \mathbb{R} \to \mathbb{R}$ by $[x]^+ = \max\{x, 0\}$ for all x in $\mathbb{R}$. Finally we define the Dirac delta function $\delta(\cdot,\cdot)$ by $\delta(i,j) = 1$ if $i = j$ and $\delta(i,j) = 0$ if $i \neq j$.

We are interested in notions of partial ordering on $\mathcal{D}(\mathbb{R}^K)$ known as *integral orderings*. Formally, with any non-empty collection Φ of Borel mappings $f : \mathbb{R}^K \to \mathbb{R}$, we associate a partial ordering $\leq_\Phi$ on $\mathcal{D}(\mathbb{R}^K)$ by saying that the RV X (with distribution F) is $\leq_\Phi$ smaller than the RV Y (with distribution G), noted $X \leq_\Phi Y$ (or equivalently $F \leq_\Phi G$), if

$$E[f(X)] \leq E[f(Y)] \tag{2.3}$$

for every mapping $f : \mathbb{R}^K \to \mathbb{R}$ in Φ, provided the expectations in (2.3) exist.

It is clear from Stoyan's monograph [43] that there are many interesting choices for Φ in this definition. In this paper, we concern ourselves with only two such choices, namely Φ_{st} and Φ_{icx} , where

$$\Phi_{st} = \{f : \mathbb{R}^K \to \mathbb{R} \text{ monotone non} - \text{decreasing}\} \tag{2.4}$$

and

$$\Phi_{icx} = \Phi_{st} \bigcap \{f : \mathbb{R}^K \to \mathbb{R} \text{ convex}\}. \tag{2.5}$$

In agreement with established usage, we substitute the notation $\leq_{st}$ (resp. $\leq_{icx}$) to $\leq_\Phi$ when $\Phi = \Phi_{st}$ (resp. $\Phi = \Phi_{icx}$) and we read $X \leq_{st} Y$ (resp. $X \leq_{icx} Y$) as saying that the RV X is smaller *stochastically* (resp. in the *convex increasing order*) than the RV Y.

The reader is referred to the monographs by Ross [8] and Stoyan [10] for additional information and properties of the orderings $\leq_{st}$ and $\leq_{icx}$. The notation used here is the one of [2] and [9].

II.2. A basic comparison result

Theorem 1 below is well known [8, Thm. 8.6.2, pp. 274-275] [10, Thm. 5.2.1, pp. 80-81] and provides a basic comparison result for recursions which arise naturally in the context of queueing systems. This result is given here for easy reference since it provides a precise mathematical underpining to the folk theorems discussed in this paper.

We start with a sequence of $\mathbb{R}$-valued RV's $\{\xi_n,\ n = 0,1,\ldots\}$, and consider the $\mathbb{R}$-valued RV's $\{W_n,\ n = 0,1,\ldots\}$ generated by the recursion

$$\begin{aligned} W_{n+1} &= [W_n + \xi_{n+1}]^+ \\ W_0 &= \xi_0. \end{aligned} \qquad n = 0,1,\ldots \tag{2.6}$$

We introduce the conditions (H1)-(H2) on the RV's $\{\xi_n,\ n = 0,1,\ldots\}$, where

(H1): The RV ξ_0 is *independent* of the sequence $\{\xi_{n+1},\ n = 0,1,\ldots\}$;

(H2): The RV's $\{\xi_{n+1},\ n = 0,1,\ldots\}$ are *i.i.d.* with common distribution F.

The familiar reader will immediately recognize in (2.6) a generalization of the Lindley recursion which describes the evolution of successive waiting times in $GI/GI/1$ queues [3,6]. For such

queueing systems, the accepted fact that "determinism minimizes waiting times" leads naturally to the question of whether the output RV's $\{W_n,\ n = 0,1,\ldots\}$ are *monotone* in the model data, namely the initial condition W and the distribution F, where monotonicity is understood in some stochastic ordering sense. More precisely, we consider the recursion (2.6) when driven by the two sequences $\{\xi_n^{(1)},\ n = 0,1,\ldots\}$ and $\{\xi_n^{(2)},\ n = 0,1,\ldots\}$, and denote their respective ouput sequences by $\{W_n^{(1)},\ n = 0,1,\ldots\}$ and $\{W_n^{(2)},\ n = 0,1,\ldots\}$.

Theorem 1. *Let $\leq$ denote either $\leq_{st}$ or $\leq_{icx}$. Assume (H1)-(H2) to hold for both driving sequences $\{\xi_n^{(1)},\ n = 0,1,\ldots\}$ and $\{\xi_n^{(2)},\ n = 0,1,\ldots\}$. With obvious meaning to the notation, if the stochastic comparisons*

$$\xi_0^{(1)} \leq \xi_0^{(2)} \quad \text{and} \quad F^{(1)} \leq F^{(2)} \tag{2.7}$$

hold, then

$$W_n^{(1)} \leq W_n^{(2)}. \qquad n = 0,1,\ldots \tag{2.8}$$

Proof. This result follows by induction on n from the fact that the mapping $x \to [x]^+$ is monotone increasing and convex, and from elementary properties of the stochastic orderings involved. Details are available in [8, Thm. 8.6.2, pp. 274-275] and in [10, Thm. 5.2.1, pp. 80-81].

□

For any sequence of $\mathbb{R}^K$-valued RV's $\{X_n,\ n = 0,1,\ldots\}$, we denote its weak limit by X_∞ (as n goes to ∞) whenever it exists, i.e., X_∞ is any $\mathbb{R}^K$-valued RV with the property that

$$P[X_\infty \leq x] = \lim_n P[X_n \leq x] \tag{2.9}$$

for all x in $\mathbb{R}^K$ which are points of continuity for the distribution of X_∞. We call X_∞ the stationary version of the sequence $\{X_n,\ n = 0,1,\ldots\}$.

Corollary. *Under the assumptions of Theorem 1, the comparison*

$$W_\infty^{(1)} \leq W_\infty^{(2)} \tag{2.10}$$

holds, provided the stationary versions of $\{W_n^{(1)},\ n = 0,1,\ldots\}$ and $\{W_n^{(2)},\ n = 0,1,\ldots\}$ exist.

Proof. The result (2.10) for $\leq_{st}$ follows from (2.8) and from the stability of $\leq_{st}$ under weak limits [10, Prop. 1.2.3, p. 6]. That (2.10) holds for $\leq_{icx}$ is a consequence of (2.8) and of the monotone character of the representation

$$W_n^{(i)} =_{st} \max\{0, \xi_1^{(i)}, \xi_1^{(i)} + \xi_2^{(i)}, \ldots, \xi_1^{(i)} + \ldots + \xi_n^{(i)}\}, \quad i = 1,2 \qquad n = 0,1,\ldots \tag{2.11}$$

for the output to (2.6). Details are available in [1,2].

□

III. PARALLEL QUEUES WITH INDEPENDENT ROUTING

In this section, we consider a queueing system composed of K (≥ 2) *identical* service stations operating in *parallel.* Each service station is constituted by a single server which is equipped with its own infinite capacity buffer and which serves customers in FCFS order. Upon arrival into the system, a customer is routed to one of the queues, with the routing decision being *independent* of the state of the system. This defines the class of *independent* assignment procedures. Of special interest to us in this class are the so-called *random* and *Round-Robin* assignments. The random assignment routes an incoming customer to the k^{th} queue with probability $\frac{1}{K}$, while the Round-Robin strategy simply assigns the n^{th} customer to the k^{th} queue if and only if $n \equiv k \pmod K$.

III.1. Parallel $GI/GI/1$ queues

In order to define a system of parallel $GI/GI/1$ queues under independent customer assignments, we start with the integrable $\mathbb{R}_+$-valued RV's $\{\tau_{n+1},\ n=0,1,\ldots\}$ and $\{\sigma_n,\ n=0,1,\ldots\}$, and with the sequence of $\{1,\ldots,K\}$-valued RV's $\{\nu_n,\ n=0,1,\ldots\}$. With this last sequence we associate a new sequence of $\{0,1\}^K$-valued RV's $\{u_n,\ n=0,1,\ldots\}$ by setting

$$u_n^k = \delta(\nu_n, k), \quad 1 \le k \le K. \qquad n=0,1\ldots(3.1)$$

These quantities are given the following interpretation: The interarrival time between the n^{th} and the $(n+1)^{rst}$ customers is given by τ_{n+1} with the convention that the 0^{th} customer arrives at time $t=0$. The n^{th} customer brings work to the system, the execution of which requires σ_n units of time, and $\nu_n = k$ (or equivalently $u_n^k = 1$) indicates that the n^{th} customer joins the k^{th} queue.

We now define the performance measures of interest under the simplifying assumption that the system is initially empty at time $t=0$. The $\mathbb{R}_+^K$-valued RV's $\{V_n,\ n=0,1,\ldots\}$ are generated componentwise by the recursion

$$\begin{aligned} V_{n+1}^k &= \left[V_n^k + u_n^k\sigma_n - \tau_{n+1}\right]^+, \quad 1 \le k \le K \\ V_0^k &= 0. \end{aligned} \qquad n=0,1,\ldots(3.2)$$

In this model, we take the view that each customer brings a task to every queue but that only the task executed by the server of the queue which the customer joins has (possibly) non-zero service duration. Clearly, V_n^k represents the work (expressed in remaining processing time) present in the k^{th} queue as the n^{th} customer enters the system, so that V_n^k is the amount of time it would have to wait in queue before receiving service if it were assigned to the k^{th} service station, i.e., if $u_n^k = 1$ The customer waiting time W_n and the system response time R_n of the n^{th} customer are thus given by

$$W_n = \sum_{k=1}^K u_n^k V_n^k \qquad n=0,1\ldots(3.3)$$

and

$$R_n = \sigma_n + \sum_{k=1}^K u_n^k V_n^k, \qquad n=0,1\ldots(3.4)$$

respectively.

Throughout the discussion we assume (A1)-(A3) to hold, namely

(A1): The three sequences $\{\tau_{n+1},\ n=0,1,\ldots\}$, $\{\sigma_n,\ n=0,1,\ldots\}$ and $\{\nu_n,\ n=0,1,\ldots\}$ are *mutually independent*;

(A2): The $\mathbb{R}_+$-valued RV's $\{\tau_{n+1},\ n=0,1,\ldots\}$ form an *i.i.d.* RV's with common distribution A;

(A3): The $\mathbb{R}_+$-valued RV's $\{\sigma_n,\ n=0,1,\ldots\}$ form an *i.i.d.* RV's with common distribution B.

The random assignment is characterized by the fact that the RV's $\{\nu_n,\ n=0,1,\ldots\}$ are *i.i.d.* RV's with common distribution given by

$$P[\nu_n = 1] = \ldots = P[\nu_n = K] = \frac{1}{K}, \qquad n=0,1\ldots(3.5)$$

whereas the Round-Robin assignment is defined by

$$\nu_n = k \quad \text{if} \quad n \equiv k \pmod K \qquad n=0,1\ldots(3.6)$$

so that the 0^{th} customer is routed to the K^{th} queue.

III.2. An alternate representation

The binary character of the assignment sequences $\{u_n^k,\ n = 0,1,\ldots\}$, $1 \le k \le K$, suggests another way of describing the evolution of the customer waiting times, namely by keeping track of the actual arrivals at a particular service station. Below we develop this alternate representation in some detail as we shall make use of it later.

For $1 \le k \le K$, we define the integer-valued RV's $\{t_m^k,\ m = 1,2,\ldots\}$ as the tags of these customers which are *effectively* routed to the k^{th} queue. Formally, we set

$$t_{m+1}^k := \begin{cases} \inf\{n > t_m^k : \nu_n = k\} & \text{if this set is not empty} \\ \infty & \text{otherwise} \end{cases} \qquad m = 0,1,\ldots (3.7)$$

with $t_0^k = 0$. To describe the behavior of the k^{th} queue as seen by the customers joining it, we define the sequences of $\mathbb{R}_+$-valued RV's $\{\tilde{\tau}_{m+1}^k,\ m = 0,1,\ldots\}$ and $\{\tilde{\sigma}_m^k,\ m = 0,1,\ldots\}$ by

$$\tilde{\tau}_{m+1}^k := \sum_{t_m^k \le n < t_{m+1}^k} \tau_{n+1} \qquad m = 0,1,\ldots (3.8)$$

and

$$\tilde{\sigma}_m^k := \sigma_{t_m^k}. \qquad m = 0,1,\ldots (3.9)$$

Note that $\tilde{\tau}_{m+1}^k$ is the time between the arrival of the m^{th} and $(m+1)^{st}$ customers to effectively enter the k^{th} queue, where the m^{th} customer receives a service of duration $\tilde{\sigma}_m^k$. From the definitions (3.8)-(3.9), we see that the waiting times $\{\tilde{W}_m^k,\ m = 1,2,\ldots\}$ for the customers joining the k^{th} queue are given simply by

$$\tilde{W}_m^k = V_{t_m^k}^k \qquad m = 1,2\ldots (3.10)$$

and that they obey the the Lindley recursion

$$\begin{aligned} \tilde{W}_{m+1}^k &= \left[\tilde{W}_m^k + \tilde{\sigma}_m - \tilde{\tau}_{m+1}\right]^+ \\ \tilde{W}_1^k &= V_{t_1^k}^k. \end{aligned} \qquad m = 1,2,\ldots (3.11)$$

If the Round-Robin assignment is used, then

$$t_m^k = mK + k, \quad 1 \le k \le K \qquad m = 1,2,\ldots (3.12)$$

Moreover, under (A1)-(A3), we see that for the random assignment, for each $1 \le k \le K$, the RV's $\{t_{m+1}^k - t_m^k,\ m = 1,2,\ldots\}$ are i.i.d RV's with a common geometric distribution given by

$$P[t_{m+1}^k - t_m^k = \ell + 1] = \frac{1}{K}\left(1 - \frac{1}{K}\right)^\ell \qquad m = 1,2,\ldots (3.13)$$

for all $l = 0,1,\ldots$.

In view of these remarks, we see that for each $1 \le k \le K$, the sequences $\{\tilde{\tau}_{m+1}^k,\ m = 0,1,\ldots\}$ and $\{\tilde{\sigma}_m^k,\ m = 0,1,\ldots\}$ are mutually independent sequences of i.i.d. RV's with common

distributions $\tilde{A}^k$ and $\tilde{B}^k$, respectively. In fact, it is a simple exercise to conclude that $\tilde{A}^1 = \ldots = \tilde{A}^K := \tilde{A}$ and $\tilde{B}^1 = \ldots = \tilde{B}^K := \tilde{B}$, with

$$\tilde{A} = \begin{cases} \sum_{\ell=0}^{\infty} \frac{1}{K}\left(1 - \frac{1}{K}\right)^{\ell} A^{*(\ell+1)} & \text{if random} \\ A^{*(K)} & \text{if Round-Robin} \end{cases} \tag{3.14}$$

and

$$\tilde{B} = B. \tag{3.15}$$

Here, for $\ell = 1, 2, \ldots$, $A^{*(\ell)}$ denotes the ℓ fold convolution of A with itself. Consequently, the output RV's $\{\tilde{W}_m^k,\ m = 1, 2, \ldots\}$ to the recursion (3.11) can be interpreted as the successive waiting times in a $GI/GI/1$ queue with interarrival time distribution $\tilde{A}$ and service time distribution $\tilde{B}$. Under both customer assignment, the stability condition of this single server queueing system is known [10, pp. 74-75] to be

$$\frac{m(\tilde{B})}{m(\tilde{A})} = \frac{1}{K} \cdot \frac{m(B)}{m(A)} < 1. \tag{3.16}$$

Theorem 2. *Assume the stability condition (3.16) to hold. Then the sequence of customer waiting times $\{W_n,\ n = 0, 1, \ldots\}$ has a stationary version W_∞. Moreover, for all $1 \leq k \leq K$, the RV's $\{\tilde{W}_m^k,\ m = 1, 2, \ldots\}$ have a stationary version $\tilde{W}_\infty$ which is independent of k, and the relation*

$$W_\infty =_{st} \tilde{W}_\infty \tag{3.17}$$

holds.

Proof. Fix $1 \leq k \leq K$. Under the condition (3.16), the sequence of RV's $\{\tilde{W}_m^k,\ m = 1, 2, \ldots\}$ has a stationary version $\tilde{W}_\infty^k$ [3,6,10] under either assignment. However, from the discussion given earlier and from (3.14)-(3.15), we see that this stationary version is independent of k [10, pp. 74-75] and we therefore denote it by $\tilde{W}_\infty$.

Under the random assignment, we have by symmetry that

$$W_n =_{st} V_n^1 =_{st} \ldots =_{st} V_n^K. \qquad n = 0, 1, \ldots \tag{3.18}$$

Since for each $1 \leq k \leq K$, the RV's $\{V_n^k,\ n = 0, 1, \ldots\}$ have a stationary version V_∞^k, so does the sequence $\{W_n,\ n = 0, 1, \ldots\}$ and (3.18) implies that

$$W_\infty =_{st} V_\infty^1 =_{st} \ldots =_{st} V_\infty^K. \qquad n = 0, 1, \ldots \tag{3.19}$$

In [1], the equality

$$V_\infty^k =_{st} \tilde{W}_\infty^k, \quad 1 \leq k \leq K \tag{3.20}$$

was established and the result thus follows.

Under the Round-Robin assignment, we see from (3.3), (3.10) and (3.12) that

$$W_n = V_{mK+k}^k = \tilde{W}_m^k \quad \text{if} \quad n = mK + k \tag{3.21}$$

for some $m = 0, 1, \ldots$ and $1 \leq k \leq K$. Therefore, for each $1 \leq k \leq K$, the RV's $\{W_{mK+k},\ m = 0, 1, \ldots\}$ have a stationary version independent of k given by the RV $\tilde{W}_\infty$ introduced earlier in the proof. Now consider a subsequence $\{W_{n_p},\ p = 0, 1, \ldots\}$ of the sequence $\{W_n,\ n = 0, 1, \ldots\}$, and observe that the index set $\{n_p,\ p = 0, 1, \ldots\}$ of this subsequence contains necessarily a subsequence

of the form $\{m_i K + \ell,\ i = 1, 2, \ldots\}$ for some $1 \leq \ell \leq K$. Consequently, the subsequence $\{W_{n_p},\ p = 0, 1, \ldots\}$ always contains a further subsequence that converges weakly (i.e., in law) and its limit is necessarily $\tilde{W}_\infty$. As a result [10, p. 153], the sequence $\{W_n,\ n = 0, 1, \ldots\}$ has a stationary version V_∞^k and (3.17) holds.

□

IV. FOLK THEOREMS FOR PARALLEL QUEUES

IV.1. A monotonicity result under the random assignment

In this section, we consider the system of parallel queues under the random assignment, and throughout the discussion, we use the superscript (K) in the notation to indicate that the quantities of interest are defined for the system with K parallel servers. We expect that as the number K of servers increases, system congestion should decrease; in particular, we expect the waiting and response times $W_n^{(K)}$ and $R_n^{(K)}$ to get smaller in some sense as K increases. We show below that this intuitive fact can be given a very precise meaning in the stochastic ordering $\leq_{st}$.

Theorem 3. *Under the assumptions (A1)-(A3), systems of parallel queues with the random customer assignment (3.5) exhibit the monotonicity properties*

$$W_n^{(K+1)} \leq_{st} W_n^{(K)} \quad \text{and} \quad R_n^{(K+1)} \leq_{st} R_n^{(K)} \qquad n = 0, 1 \ldots (4.1)$$

for all $K \geq 1$. A stationary version of (4.1) holds whenever appropriate.

Note that elementary coupling arguments can be given to compare the system with K servers to the one with $K \cdot L$ servers. The basis for this approach lies in the fact that the random assignment can be implemented in two steps by fictitiously grouping the $K \cdot L$ servers into K groups of L servers each. Indeed it is equivalent to first select randomly one of the K groups of servers (with probability $\frac{1}{L}$) and then to choose at random a server within the selected group (with probability $\frac{1}{K}$). Unfortunately, this elementary approach does not seem to allow for the comparison between the systems with K and $K+1$ servers, and a more analytical proof based on Theorem 1 is required.

Proof. The $\{0, 1\}$-valued RV's $\{b_n^{(K)},\ n = 0, 1, \ldots\}$ defined by

$$b_n^{(K)} = \delta(\nu_n^{(K)}, 1) \qquad n = 0, 1 \ldots (4.2)$$

form an i.i.d. sequence with

$$P[b_n^{(K)} = 1] = 1 - P[b_n^{(K)} = 0] = \frac{1}{K}. \qquad n = 0, 1 \ldots (4.3)$$

Moreover, we see from (A1)-(A3) that the $\mathbb{R}_+$-valued RV's $\{U_n^{(K)},\ n = 0, 1, \ldots\}$ generated by the Lindley recursion

$$\begin{aligned} U_{n+1}^{(K)} &= \left[U_n^{(K)} + b_n^{(K)} \sigma_n - \tau_{n+1} \right]^+ \\ U_0^{(K)} &= 0. \end{aligned} \qquad n = 0, 1 \ldots (4.4)$$

are the succesive customer waiting times in a $GI/GI/1$ queue with interarrival times $\{\tau_{n+1},\ n = 0, 1, \ldots\}$ and service times $\{b_n^{(K)} \sigma_n,\ n = 0, 1, \ldots\}$.

Fix $K = 1, 2, \ldots$. From (3.2)-(3.3), we see by symmetry that

$$W_n^{(K)} =_{st} U_n^{(K)}. \qquad n = 0, 1 \ldots (4.5)$$

Under (A1), the RV's $b_n^{(K)}$ and σ_n are independent, so that

$$P[b_n^{(K)}\sigma_n > t] = \frac{1}{K}[1 - B(t)]. \quad t \geq 0. \qquad n = 0, 1 \ldots (4.6)$$

We then conclude [8, Prop. 8.1.2, p. 252] that

$$b_n^{(K+1)}\sigma_n \leq_{st} b_n^{(K)}\sigma_n \qquad n = 0, 1 \ldots (4.7)$$

and therefore

$$b_n^{(K+1)}\sigma_n - \tau_{n+1} \leq_{st} b_n^{(K)}\sigma_n - \tau_{n+1} \qquad n = 0, 1 \ldots (4.8)$$

under the independence assumption (A1) [8, p. 256]. Since $U_0^{(K)} = U_0^{(K+1)} = 0$, all the conditions are in place for applying Theorem 1 to the recursion (4.4) and we obtain

$$U_n^{(K+1)} \leq_{st} U_n^{(K)}. \qquad n = 0, 1 \ldots (4.9)$$

The first inequality in (4.1) now follows upon combining (4.4) and (4.9); the second inequality is now immediate since the RV σ_n is independent of both $W_n^{(K)}$ and $W_n^{(K+1)}$ [8, p. 256].

□

IV.2. Round-Robin vs. Random

We now compare the system of parallel queues under random customer assignment against the same system under the Round-Robin assignment. It is intuitively perceived - and widely accepted - that the Round-Robin assignment outperforms the random assignment since the effective arrival stream to individual queues exhibits less statistical variability under the former assignment. This folk result is often demonstrated when the input stream and the service times are exponentially distributed for in that case *explicit* formulae are available for the mean waiting and response times in steady state under both customer assignments. Indeed, in statistical equilibrium, the system then behaves under the random assignment like an $M|M|1$ queue, while under the Round-Robin assignment it behaves like an $E_K|M|1$ queue.

We show here that this folk result can be given a precise formulation in the ordering $\leq_{icx}$; the validity of the statement is more general than the one usually made on the mean values, and is independent of any distributional asssumptions on the arrival and service processes.

Theorem 4. *Assume the stability condition (3.16) to hold. With an obvious meaning to the notation, the stochastic comparison*

$$W_\infty^{RR} \leq_{icx} W_\infty^{ran} \qquad (4.10)$$

holds.

It should be pointed out that (4.10) does not hold in the transient regime nor does it hold in the stronger ordering $\leq_{st}$, as can be seen from simple counter-examples. The proof of this result is based on the following lemma which is a particular case of Lemma 8.6.7 in [8, pp. 278-279].

Lemma 5. *Let $\{X_m, m = 1, 2, \ldots\}$ be a sequence of i.i.d. $\mathbb{R}_+$-valued RV's with common distribution F, and define the sequence of partial sums $\{S_m, m = 0, 1, \ldots\}$ by $S_m := \sum_{k=1}^m X_m$ for all $m = 1, 2, \ldots$ with $S_0 = 0$. Let α and β be integrable $\{m = 0, 1, \ldots\}$-valued RV's which are each independent of the sequence $\{X_m, m = 1, 2, \ldots\}$. The stochastic comparison $\alpha \leq_{icx} \beta$ then implies*

$$S_\alpha \leq_{icx} S_\beta. \qquad (4.11)$$

Proof. For any Borel mapping $f : \mathbb{R}_+ \to \mathbb{R}$, we use the assumed independence to write

$$E[f(S_\alpha)] = \sum_{m=0}^{\infty} P[\alpha = m]\hat{f}(m) \tag{4.12}$$

where

$$\hat{f}(m) := E[f(S_m)]. \qquad m = 0, 1, \ldots (4.13)$$

provided these expectations exist; an expression similar to (4.12) is available for $E[f(S_\beta)]$.

Whenever $f : \mathbb{R}_+ \to \mathbb{R}$ is monotone increasing, we see that $m \to \hat{f}(m)$ is also monotone increasing since $X_{m+1} \geq 0$, whence $S_m \leq S_{m+1}$, for all $m = 0, 1, \ldots$. Moreover, if $f : \mathbb{R}_+ \to \mathbb{R}$ is convex, then we readily see that

$$E[f(S_{m-1} + X_m) - f(S_{m-1})] \leq E[f(S_{m-1} + X_m + X_{m+1}) - f(S_{m-1} + X_m)]$$

$$m = 1, 2, \ldots (4.14)$$

since the non-negative RV's $\{X_m, m = 0, 1, \ldots\}$ are i.i.d, and the mapping $m \to \hat{f}(m)$ is thus integer-convex, i.e.,

$$\hat{f}(m) - \hat{f}(m-1) \leq \hat{f}(m+1) - \hat{f}(m). \qquad m = 1, 2, \ldots (4.15)$$

The linear interpolation of the sequence $\{\hat{f}(m),\ m = 0, 1, \ldots\}$ is the mapping $\hat{f}_c : \mathbb{R}_+ \to \mathbb{R}$ defined by

$$\hat{f}_c(t) := \hat{f}(m) + [\hat{f}(m+1) - \hat{f}(m)](t - m) \quad \text{if} \quad m \leq t \leq m + 1. \qquad m = 0, 1, \ldots (4.16)$$

With this notation, we can write (4.12) as

$$E[f(S_\alpha)] = E[\hat{f}_c(\alpha)] \quad \text{and} \quad E[f(S_\beta)] = E[\hat{f}_c(\beta)]. \tag{4.17}$$

Whenever the mapping $f : \mathbb{R}_+ \to \mathbb{R}$ is monotone increasing and convex, so is $\hat{f}_c : \mathbb{R}_+ \to \mathbb{R}$ by virtue of the remarks made earlier on the sequence $\{\hat{f}(m),\ m = 0, 1, \ldots\}$. The assumption $\alpha \leq_{icx} \beta$ thus translates into $E[\hat{f}_c(\beta)] \leq E[\hat{f}_c(\alpha)]$, or equivalently into $E[f(S_\alpha)] \leq E[f(S_\beta)]$ upon making use of (4.17). This complete the proof.

□

The following special case of Lemma 5 is of use in the sequel. If the RV's α and β are chosen such that $\alpha = K$ and $E[\beta] = K$, then $K \leq_{icx} \beta$ by Jensen's inequality and therefore $S_K \leq_{icx} S_\beta$ by invoking Lemma 5.

A proof of Theorem 4. Having in mind the alternate representations developed in Section III, we consider two auxiliary Lindley recursions, namely

$$\tilde{W}_{n+1}^{RR} = \left[\tilde{W}_n^{RR} + \sigma_n - \tau_{n+1}^{RR}\right]^+ \qquad n = 0, 1, \ldots (4.18)$$
$$\tilde{W}_0^{RR} = 0$$

and

$$\tilde{W}_{n+1}^{ran} = \left[\tilde{W}_n^{ran} + \sigma_n - \tau_{n+1}^{ran}\right]^+ \qquad n = 0, 1, \ldots (4.19)$$
$$\tilde{W}_0^{ran} = 0.$$

Each one of the sequences $\{\tau_{n+1}^{RR},\ n = 0,1,\ldots\}$ and $\{\tau_{n+1}^{ran},\ n = 0,1,\ldots\}$ is a sequence of i.i.d $I\!R_+$-valued RV's with common distribution A^{RR} and A^{ran} respectively given by

$$A^{RR} = A^{*(K)} \quad \text{and} \quad A^{ran} = \sum_{\ell=0}^{\infty} \frac{1}{K}\left(1 - \frac{1}{K}\right)^{\ell} A^{*(\ell+1)}. \tag{4.20}$$

Moreover, each one of these sequences is assumed independent of the sequence of i.i.d RV's $\{\sigma_n,\ n = 0,1,\ldots\}$ the common distribution of which is simply B.

Under the stability condition (3.16), the sequences $\{\tilde{W}_n^{RR},\ n = 0,1,\ldots\}$ and $\{\tilde{W}_n^{ran},\ n = 0,1,\ldots\}$ have stationary versions, say $\tilde{W}_\infty^{RR}$ and $\tilde{W}_\infty^{ran}$. A careful inspection of the proof of Theorem 2 shows that

$$\tilde{W}_\infty^{RR} =_{st} W_\infty^{RR} \quad \text{and} \quad \tilde{W}_\infty^{ran} =_{st} W_\infty^{ran}. \tag{4.21}$$

With the notation of Lemma 5, we see that $A^{RR} =_{st} S_K$ and that $A^{ran} =_{st} S_\beta$ where β is an $\{1,2,\ldots\}$-valued RV with geometric distribution given by

$$P[\beta = \ell + 1] = \frac{1}{K}\left(1 - \frac{1}{K}\right)^{\ell} \qquad \ell = 0,1,\ldots \tag{4.22}$$

and the i.i.d RV's $\{X_m,\ m = 0,1,\ldots\}$ have common distribution A.

By the remark following Lemma 5, we have $A^{RR} \leq_{icx} A^{ran}$ or equivalently that

$$\tau_{n+1}^{RR} \leq_{icx} \tau_{n+1}^{ran}. \qquad n = 0,1,\ldots \tag{4.23}$$

Since $m(A^{RR}) = m(A^{ran}) = Km(A)$, we conclude from [8, Cor. 8.5.3, p. 272]

$$-\tau_{n+1}^{RR} \leq_{icx} -\tau_{n+1}^{ran} \qquad n = 0,1,\ldots \tag{4.24}$$

and therefore

$$\sigma_n - \tau_{n+1}^{RR} \leq_{icx} \sigma_n - \tau_{n+1}^{ran} \qquad n = 0,1,\ldots \tag{4.25}$$

upon invoking the independence of the arrival and service sequences [8, Prop. 8.5.4, p. 272]. A straightforward application of the corollary to Theorem 1 now implies the comparison

$$\tilde{W}_\infty^{RR} \leq_{icx} \tilde{W}_\infty^{ran} \tag{4.26}$$

which is equivalent to (4.10) by virtue of (4.21).

□

V. CONCLUSIONS

By an appropriate choice of representation, we have been able to provide simple proofs of some well-known folk theorems of queueing theory. In particular, we have shown that the response time of a multi-server queue with random routing is stochastically decreasing in the number of servers, and that the equilibrium response time in a system with Round-Robin customer assignment is smaller (in the convex increasing order) than in an identical system with random routing. We believe that the techniques developed in this paper can be extended to allow the formalization of many such folk theorems.

REFERENCES

[1] F. Baccelli, A. M. Makowski and A. Shwartz, "The Fork-Join queue and related systems with synchronization constraints: Stochastic ordering and computable bounds", To appear in *Advances in Applied Probability*, Also available as Technical Research Report **TR-87-01**, Systems Research Center, University of Maryland, College Park, Maryland, and Rapport de Recherche **687**, INRIA-Rocquencourt (France), June 1987.

[2] F. Baccelli and A. M. Makowski, "Queueing models for systems with synchronization constraints," *Proceedings of the IEEE* **77** (1989), Invited paper, Special Issue on Dynamics of Discrete Event Systems, pp. 138-161.

[3] J.W. Cohen, *The Single Server Queue*, North-Holland, Amsterdam (The Netherlands), 1969.

[4] B. Hajek, "The proof of a folk theorem on queueing delay with applications to routing in networks," *J. Assoc. Comp. Mach.* **30** (1983), pp. 834-851.

[5] P. Humblet, *Determinism Minimizes Waiting Times in Queues,* Technical Report, LIDS - Department of Electrical Engineering and Computer Science, M.I.T, Cambridge (MA), 1982.

[6] L. Kleinrock, *Queueing Systems I: Theory*, J. Wiley & Sons, New York (NY), 1976.

[7] B.A. Rogozin, "Some extremal problems in queueing theory," *Theor. Prob. Appl.* **11** (1966), pp. 144-151.

[8] S. Ross, *Stochastic Processes*, J. Wiley & Sons, New York (NY), 1984.

[9] M. Shaked and J.G. Shantikumar, "Stochastic convexity and its applications," *Advances in Applied Probability* **20** (1988), pp. 427-446.

[10] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, English Translation (D.J. Daley, Editor), J. Wiley & Sons, New York (NY), 1984.

[11] W. Whitt, "Minimizing delays in the GI/GI/1 queue," *Opns. Res.* **32** (1984), pp. 41-51.