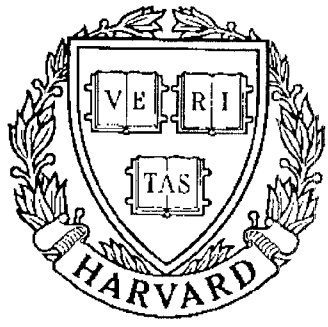
TECHNICAL RESEARCH REPORT



SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Convexity Results for Parallel Queues with Bernoulli Routing

by L. Gün, A. Jean-Marie, A.M. Makowski and Tedijanto

TR 90-52

CONVEXITY RESULTS FOR PARALLEL QUEUES WITH BERNOULLI ROUTING

Levent Gün[1]

IBM – Network Analysis Center, P.O. Box 12195 E95/B673, Research Triangle Park, Durham, NC 27709.

Alain Jean–Marie [2]

INRIA – Sophia Antipolis, 2004 Route des Lucioles, 06565 Valbonne Cédex, France.

Armand M. Makowski [3]

Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, MD 20742.

and

Tedijanto [4]

IBM – Network Analysis Center, P.O. Box 12195 E95/B673, Research Triangle Park, Durham, NC 27709.

ABSTRACT

In this note, we derive various convexity and monotonicity properties of performance measures in parallel $GI/GI/1$ queues with Bernoulli routing. We then use these results to establish that the equilikely assignment is optimal among all Bernoulli routing. The main tools are notions of stochastic convexity recently introduced by Shaked and Shanthikumar. The optimality results are couched in terms of the stochastic orderings $\leq_{st}$ and $\leq_{icx}$.

Key Words: Bernoulli routing, stochastic convexity, Lindley recursion, equilikely assignment

[1] The work of this author was supported partially through NSF Grant ECS-83-51836 and partially through a grant from AT&T Bell Laboratories.

[2] The work of this author was performed while on leave from INRIA.

[3] The work of this author was supported partially through NSF Grant ECS-83-51836 and partially through NSF Engineering Research Centers Program Grant CDR-88-03012-04.

[4] The work of this author was supported through NSF Engineering Research Centers Program Grant CDR-88-03012-04.

I. INTRODUCTION

Roughly speaking, a Bernoulli mechanism is one whereby decisions are generated through a sequence of independent identically distributed (i.i.d.) random variables (rvs) which take on only finitely many values. In queueing systems such mechanisms are often used to implement decisions variables in the absence of state information, in which case the decision is taken independently of any event, past, present or future. Typical examples include random routing in BCMP networks, Bernoulli schedules in vacation models and randomized priorities for server assignment in multi–class systems.

In this note, we are concerned with the convexity properties of several performance measures in parallel $GI/GI/1$ queues with Bernoulli or random routing. We also derive several optimality results for the equilikely assignment; these optimality results are couched in terms of the stochastic orderings $\leq_{st}$ and $\leq_{icx}$ (Ross [10], Stoyan [13]). Our main tools are the notions of stochastic convexity recently introduced by Shaked and Shanthikumar ([11,12]), and their properties.

The paper is organized as follows: In Section 2, we summarize the notions of stochastic convexity which we use in the discussion. A basic convexity result for Lindley recursions is then presented in Section 3. This is followed by Section 4 where a model of parallel $GI/GI/1$ queues with Bernoulli routing is introduced, and preliminary convexity results are derived for customer waiting and response times. In Section 5, we make use of these results to show that the waiting and response time of any customer are each minimized by the equilikely assignment in the sense of the ordering $\leq_{icx}$. Section 6 is devoted to the derivation of a similar result for a system of parallel $M/GI/1$ queues; the optimality of the equilikely assignment is obtained in the sense of the stronger ordering $\leq_{st}$ but only in statistical equilibrium. We conclude with Section 7 which contains several remarks on the obtained results.

The notation adopted here is the one used in (Baccelli and Makowski [1]) and (Shaked and Shanthikumar [11,12]): We denote the set of real (resp. non–negative real) numbers by $\mathbb{R}$ (resp. $\mathbb{R}_+$). The kth component of any element x in $\mathbb{R}^K$ is denoted by x^k, $1 \leq k \leq K$. 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 then said to be increasing (resp. decreasing) if $x \leq y$ in $\mathbb{R}^K$ implies $f(x) \leq f(y)$ (resp. $f(x) \geq f(y)$).

We find it convenient to define all the random variables of interest on some common probability triple $(\Omega, \mathcal{F}, P)$. We identify a probability distribution function F on $\mathbb{R}$ with an $\mathbb{R}$–valued rv X which has distribution F, in which case $F(x) = P[X \leq x]$ for all x in $\mathbb{R}$, and we set

$$m(F) = \int_{\mathbb{R}} x dF(x) = m(X) \tag{1.1}$$

whenever the first moment of F exists. Finally two $\mathbb{R}$–valued rvs X and Y are said to be *equal in law* if they have the same distribution, a fact we denote by $X =_{st} Y$.

The reader is referred to the monographs by Ross [10] and Stoyan [14] for additional information and properties of the stocastic orderings $\leq_{st}$ and $\leq_{icx}$.

II. PRELIMINARIES ON STOCHASTIC CONVEXITY

In this section, we briefly recall several notions of stochastic convexity which have been recently introduced by Shaked and Shanthikumar [11,12]: Throughout, Θ is a convex subset of $\mathbb{R}$ and $\{X(\theta),\ \theta \in \Theta\}$ is a collection of $\mathbb{R}$–valued rvs. For any Borel mapping $f : \mathbb{R} \to \mathbb{R}$, we define the mapping $\hat{f} : \Theta \to \mathbb{R}$ by

$$\hat{f}(\theta) := E[f(X(\theta))], \quad \theta \in \Theta \tag{2.1}$$

whenever these expectations exist. The collection of rvs $\{X(\theta),\ \theta \in \Theta\}$ is then said to be

1. stochastically increasing (resp. decreasing) convex in the usual stochastic ordering – in short SICX(st) (resp. SDCX(st)) – if $\hat{f}$ is increasing (resp. decreasing) convex whenever f is increasing;

2. stochastically increasing (resp. decreasing) convex – in short SICX (resp. SDCX) – if $\hat{f}$ is increasing (resp. decreasing) whenever f is increasing *and* $\hat{f}$ is increasing (resp. decreasing) convex whenever f is increasing convex;

3. stochastically increasing convex in the sample path sense – in short SICX(sp) – if for any four points θ_i, $i = 1, \ldots, 4$, in Θ, such that $\theta_1 \leq \theta_2 \leq \theta_3 \leq \theta_4$ and $\theta_1 + \theta_4 = \theta_2 + \theta_3$, there exist four rvs X_i, $i = 1, \ldots, 4$, defined on a common probability space such that $X_i =_{st} X(\theta_i)$, $i = 1, \ldots, 4$, and

$$X_j \leq X_4, \quad j = 1, 2, 3 \quad \text{and} \quad X_2 + X_3 \leq X_1 + X_4 \quad a.s. \tag{2.2}$$

A few words on these definitions: When the rvs $\{X(\theta),\ \theta \in \Theta\}$ are non–negative rvs, we note in the definition of SICX(st) and SICX that we need only consider $\mathbb{R}_+$–valued mappings f, in which case $\hat{f}$ is always well defined. Moreover, when Θ is a subset of $\{0, 1, \ldots\}$, convexity is understood as integer convexity.

The following implications were discussed in (Shaked and Shanthikumar [11,12]):

$$\text{SICX(st)} \Longrightarrow \text{SICX(sp)} \Longrightarrow \text{SICX}. \tag{2.3}$$

In general, the implications SICX(sp) $\Longrightarrow$ SICX(st) and SICX $\Longrightarrow$ SICX(sp) are not true as can be seen on simple counterexamples.

III. A BASIC CONVEXITY RESULT

Theorem 1 below is borrowed from (Shaked and Shanthikumar [11, Thm. 5.1, p. 439]), and provides a basic convexity result for recursions which arise naturally in the context of queueing systems. We give this result in a form suitable for easy reference.

Let Θ be a convex subset of $\mathbb{R}$. For each θ in Θ, we start with a sequence of $\mathbb{R}$–valued rvs $\{\xi_{n+1}(\theta),\ n = 0, 1, \ldots\}$, and consider the $\mathbb{R}$–valued rvs $\{W_n(\theta),\ n = 0, 1, \ldots\}$ generated by the recursion

$$\begin{aligned} W_{n+1}(\theta) &= \left[W_n(\theta) + \xi_{n+1}(\theta)\right]^+ \\ W_0(\theta) &= 0 \end{aligned} \qquad n = 0, 1, \ldots \tag{3.1}$$

with the notation $x^+ = \max\{0, x\}$ for all x in $\mathbb{R}$. We introduce the condition **(H1)** on the sequence driving the recursion (3.1).

(H1): For each θ in Θ, the rvs $\{\xi_{n+1}(\theta),\ n = 0, 1, \ldots\}$ are *i.i.d.* with common distribution F_θ. Throughout $\xi(\theta)$ denotes a generic rv distributed according to F_θ.

The familiar reader will immediately recognize in (3.1) a generalization of the Lindley recursion which describes the evolution of successive waiting times in $GI/GI/1$ queues (Cohen [5], Kleinrock [8]).

Theorem 1. *Assume* **(H1)** *to hold. If* $\{\xi(\theta),\ \theta \in \Theta\}$ *is SICX(sp), then for each* $n = 0, 1, \ldots$, $\{W_n(\theta),\ \theta \in \Theta\}$ *is SICX(sp).*

Proof. This result follows by induction on n from the fact that the mapping $x \to x^+$ is increasing convex, and from elementary properties of SICX(sp). Details are available in (Shaked and Shanthikumar [11, p. 439]).

■

By combining Theorem 1 with remark (2.3), we obtain

Corollary 2. *Under the assumptions of Theorem 1, for each* $n = 0, 1, \ldots$, $\{W_n(\theta),\ \theta \in \Theta\}$ *is SICX, i.e., for every increasing convex mapping* $f : \mathbb{R}_+ \to \mathbb{R}$, *the mapping* $\hat{f}_n : \Theta \to \mathbb{R}$ *defined by*

$$\hat{f}_n(\theta) := E[f(W_n(\theta))], \quad \theta \in \Theta \tag{3.2}$$

is also increasing convex.

For any sequence of $\mathbb{R}^K$–valued rvs $\{X_n,\ n = 0, 1, \ldots\}$, 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

$$P[X_\infty \leq x] = \lim_n P[X_n \leq x] \tag{3.3}$$

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 $\{X_n,\ n = 0, 1, \ldots\}$. It is well known that under **(H1)**, for each θ in Θ the sequence $\{W_n(\theta),\ n = 0, 1, \ldots\}$ admits a stationary version if and only

$$m(F_\theta) < 0 \tag{3.4}$$

provided of course that the first moment of F_θ exists and is finite.

Corollary 3. *Under the assumptions of Theorem 1 with (3.4) enforced for each* θ *in* Θ, $\{W_\infty(\theta),\ \theta \in \Theta\}$ *is SICX.*

Proof. For each θ in Θ, we note the monotone character of the representation

$$W_n(\theta) =_{st} \max\{0, \xi_1(\theta), \xi_1(\theta) + \xi_2(\theta), \ldots, \xi_1(\theta) + \ldots + \xi_n(\theta)\} \qquad n = 0, 1, \ldots \tag{3.5}$$

for the output to (3.1). By the Monotone Convergence Theorem we then conclude that

$$\lim_n E[f(W_n(\theta))] = E[f(W_\infty(\theta))], \quad \theta \in \Theta \tag{3.6}$$

for every increasing convex mapping $f : \mathbb{R} \to \mathbb{R}_+$. The result now follows from Corollary 2 and from the fact that the pointwise limit of a sequence of increasing convex mappings is also increasing convex.

■

IV. PARALLEL QUEUES WITH BERNOULLI ROUTING

In this section, we consider a queueing system composed of K (≥ 2) 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 according to the FCFS discipline. Upon arrival into the system, a customer is routed to one of the queues, with the routing decision being *independent* of any other event, past, present or future. An incoming customer is routed to the kth queue with probability q^k, $1 \leq k \leq K$. The probability vector $q = (q^1, \ldots, q^K)$ is an element of the simplex $\mathcal{S}_K$ defined by

$$\mathcal{S}_K := \{x \in [0,1]^K : \sum_{k=1}^{K} x^k = 1\}. \tag{4.1}$$

In order to define such a system of parallel $GI/GI/1$ queues under this random customer assignment, we start with the integrable $\mathbb{R}_+$–valued rvs $\{\tau_{n+1},\ n = 0,1,\ldots\}$, the integrable $\mathbb{R}_+^K$–valued rvs $\{\sigma_n = (\sigma_n^1, \ldots, \sigma_n^K),\ n = 0,1,\ldots\}$, and the $\{1,\ldots,K\}$–valued rvs $\{\nu_n(q),\ n = 0,1,\ldots\}$. With this last sequence we associate a new sequence of $\{0,1\}^K$–valued rvs $\{u_n(q),\ n = 0,1,\ldots\}$ by setting

$$u_n^k(q) = \delta(\nu_n(q), k), \quad 1 \leq k \leq K. \qquad n = 0,1\ldots \tag{4.2}$$

In (4.2) we have used the notation $\delta(i,j) = 1$ if $i = j$ and $\delta(i,j) = 0$ if $i \neq j$, $1 \leq i,j \leq K$.

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

We define the performance measures of interest under the simplifying assumption that the system is initially empty at time $t = 0$. For $1 \leq k \leq K$, we generate the $\mathbb{R}_+$–valued rvs $\{V_n^k(q),\ n = 0,1,\ldots\}$ by the recursion

$$\begin{aligned} V_{n+1}^k(q) &= \left[V_n^k(q) + u_n^k(q)\sigma_n^k - \tau_{n+1}\right]^+ \qquad n = 0,1,\ldots \\ V_0^k(q) &= 0. \end{aligned} \tag{4.3}$$

In this formulation, 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, $V_n^k(q)$ represents the work (expressed in

remaining processing time) present in the kth queue as the nth customer enters the system. In other words, $V_n^k(q)$ is the amount of time it would have to wait in queue before receiving service if it were assigned to the kth service station, i.e., if $u_n^k(q) = 1$ The waiting time $W_n(q)$ and the system response time $R_n(q)$ of the nth customer are thus given by

$$W_n(q) = \sum_{k=1}^{K} u_n^k(q) V_n^k(q) \qquad n = 0, 1 \ldots (4.4)$$

and

$$R_n(q) = \sum_{k=1}^{K} u_n^k(q) \cdot \left(V_n^k(q) + \sigma_n^k\right), \qquad n = 0, 1 \ldots (4.5)$$

respectively.

Throughout the discussion we assume conditions **(A1)**–**(A4)** to hold, namely

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

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

(A3): The $\mathbb{R}_+^K$–valued rvs $\{\sigma_n,\ n = 0, 1, \ldots\}$ form an *i.i.d.* sequence. Moreover, for each $1 \leq k \leq K$, the i.i.d rvs $\{\sigma_n^k,\ n = 0, 1, \ldots\}$ have common distribution B_k; and

(A4): The $\{1, \ldots, K\}$–valued rvs $\{\nu_n(q),\ n = 0, 1, \ldots\}$ form an *i.i.d.* sequence with common distribution

$$P[\nu_n(q) = k] = q^k, \quad 1 \leq k \leq K. \qquad n = 0, 1 \ldots (4.6)$$

Fix q in $\mathcal{S}_K$ and observe from **(A4)** that for each $1 \leq k \leq K$, the rvs $\{u_n^k(q),\ n = 0, 1, \ldots\}$ are i.i.d. Bernoulli rvs with parameter q^k, i.e.,

$$P[u_n^k(q) = 1] = q^k = 1 - P[u_n^k(q) = 0]. \qquad n = 0, 1, \ldots (4.7)$$

Consequently the statistics of the rvs $\{u_n^k(q),\ n = 0, 1, \ldots\}$ are fully specified by q^k, and without any danger of confusion, we can write $\{u_n^k(q^k),\ n = 0, 1, \ldots\}$ in lieu of $\{u_n^k(q),\ n = 0, 1, \ldots\}$ as we do from now on. As a result, the statistics of the rvs $\{V_n^k(q),\ n = 0, 1, \ldots\}$ depend on q only through q^k, and again we shall use the simpler notation $\{V_n^k(q^k),\ n = 0, 1, \ldots\}$ from now on.

We are now ready to state the main result of this section.

Theorem 4. *Assume* **(A1)**–**(A4)** *to hold for each q in $\mathcal{S}_K$. For each $n = 0, 1, \ldots$, the following statements hold true.*

1. *For each $1 \leq k \leq K$, $\{V_n^k(q^k),\ q^k \in [0, 1]\}$ is SICX(sp), thus SICX;*
2. *For each $1 \leq k \leq K$, $\{V_n^k(q^k) + \sigma_n^k,\ q^k \in [0, 1]\}$ is SICX(sp), thus SICX; and*
3. *For every increasing convex mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$, the mappings $q \to E[f(W_n(q))]$ and $q \to E[f(R_n(q))]$ are convex on $\mathcal{S}_K$.*

Proof. Fix $1 \leq k \leq K$. For each q^k in $[0,1]$, define the rvs $\{\xi_{n+1}(q^k),\ n = 0,1,\ldots\}$ by

$$\xi_{n+1}(q^k) = u_n^k(q^k)\sigma_n^k - \tau_{n+1}. \qquad n = 0,1,\ldots \quad (4.8)$$

As pointed out in (Shaked and Shanthikumar [11, Ex. 4.4, p. 438]), for each $n = 0,1,\ldots$, $\{u_n^k(q^k),\ q^k \in [0,1]\}$ is SICX(sp) whence $\{\xi_{n+1}(q^k),\ q^k \in [0,1]\}$ is also SICX(sp) by making use of the independence assumption **(A1)** (Shaked and Shanthikumar [11, Thm. 3.10, p. 436]). Claim 1 is now immediate from Theorem 1 (and Corollary 2) since **(A1)**–**(A4)** imply **(H1)**.

Next, we observe for each $n = 0,1,\ldots$, that the rvs $V_n^k(q^k)$ and σ_n^k are independent under **(A1)**–**(A4)**. Since $\{V_n^k(q^k),\ q^k \in [0,1]\}$ is SICX(sp) by Claim 1, we readily conclude that $\{V_n^k(q^k)+\sigma_n^k,\ q^k \in [0,1]\}$ is also SICX(sp) (Shaked and Shanthikumar [11, Thm. 3.10, p. 436]). We finally complete the proof of Claim 2 by invoking Corollary 2.

Now, we fix $n = 0,1,\ldots$ and q in $\mathcal{S}_K$. For every Borel mapping $f : I\!R_+ \to I\!R_+$, we conclude from (4.4) that

$$\begin{aligned} E[f(W_n(q))] &= E\left[\sum_{k=1}^{K} u_n^k(q) f(V_n^k(q^k))\right] \\ &= \sum_{k=1}^{K} q^k E[f(V_n^k(q^k))]. \end{aligned} \quad (4.9)$$

The second equality follows from the fact that for each $1 \leq k \leq K$, the rvs $u_n^k(q^k)$ and $V_n^k(q^k)$ are independent by virtue of **(A1)**–**(A4)**. This last relation can be rewritten as

$$E[f(W_n(q))] = \sum_{k=1}^{K} q^k \hat{f}_n^k(q^k), \quad q \in \mathcal{S}_K \quad (4.10)$$

where for each $1 \leq k \leq K$, we define the mapping $\hat{f}_n^k : [0,1] \to I\!R_+$ by

$$\hat{f}_n^k(q^k) := E[f(V_n^k(q^k))], \quad q^k \in [0,1]. \quad (4.11)$$

If f is assumed increasing convex, then for each $1 \leq k \leq K$, the mapping $\hat{f}_n^k : [0,1] \to I\!R_+$ is also increasing convex by Claim 1 (and Corollary 2), and so is the mapping $q^k \to q^k \hat{f}_n^k(q^k)$. With the help of these facts, it is now a simple exercise to conclude from (4.10) to the asserted convexity of the mapping $q \to E[f(W_n(q))]$.

The convexity of the mapping $q \to E[f(R_n(q))]$ is established along similar lines, with (4.5) and Claim 2 as points of departure; details are omitted in the interest of brevity. ∎

It is worth pointing out that for every mapping $f : I\!R_+ \to I\!R_+$, the mapping $q \to \sum_{k=1}^{K} q^k \hat{f}_n^k(q^k)$ is always well defined on the *entire* K–dimensional cube $[0,1]^K$, and not merely on the simplex $\mathcal{S}_K$. The arguments given in the proof of Theorem 4 show that

this mapping is increasing convex on the entire set $[0,1]^K$ whenever f is increasing convex. However, it is meaningless to consider the monotonicity of this function on the simplex $\mathcal{S}_K$ since there the componentwise ordering reduces to equality (i.e., for all q and q' in $\mathcal{S}_K$, $q \leq q'$ implies $q = q'$). The convexity of this extended function can nevertheless be used to advantage in designing computational algorithms to minimize any performance measure of the form $E(f(W_n(q)))$ with an increasing convex mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$. Similar comments apply when considering response times.

V. THE HOMOGENEOUS CASE

Of special interest is the *homogeneous* situation in which case all processors are identical with

$$B_1 = \ldots = B_K = B. \tag{5.1}$$

As we shall now see, in the homogeneous situation, Theorem 4 has several interesting implications which are presented in Theorems 5 and 6 below.

Throughout we denote by e_K the vector of ones in $\mathbb{R}^K$, i.e., $e_K = (1, \ldots, 1)$.

Theorem 5. *Assume* **(A1)–(A4)** *to hold for each q in $\mathcal{S}_K$. Under the homogeneity assumption (5.1), the stochastic ordering relations*

$$W_n(\frac{1}{K}e_K) \leq_{icx} W_n(q), \quad q \in \mathcal{S}_K \qquad n = 0, 1, \ldots \tag{5.2}$$

and

$$R_n(\frac{1}{K}e_K) \leq_{icx} R_n(q), \quad q \in \mathcal{S}_K \qquad n = 0, 1, \ldots \tag{5.3}$$

hold true, i.e., the probability vector $q^* = \frac{1}{K}e_K$ minimizes the waiting and response times of the nth customer in the sense of the ordering $\leq_{icx}$.

Elements of a proof of this result have been independently discussed by Chang, Chao and Pinedo [3,4]. Tsoucas [14] has recently obtained a related result for systems of parallel $M/M/1$ queues. Among other things, he showed by coupling arguments that at all times, the total number of customers in the system is minimized by $q^* = \frac{1}{K}e_K$ in the sense of the ordering $\leq_{st}$.

Proof. Fix $n = 0, 1, \ldots$. Under the homogeneity assumption (5.1), for every Borel mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$, we observe (with the notation (4.11)) that

$$\hat{f}_n^1(p) = \ldots = \hat{f}_n^K(p) =: \hat{f}_n(p), \quad p \in [0,1] \tag{5.4}$$

so that (4.10) becomes

$$E[f(W_n(q))] = \sum_{k=1}^{K} q^k \hat{f}_n(q^k), \quad q \in \mathcal{S}_K. \tag{5.5}$$

Now if f is assumed increasing convex, then the mapping $\hat{f}_n : [0,1] \to \mathbb{R}_+$ is also increasing convex by Claim 1 of Theorem 4. Therefore, the convexity of $\hat{f}_n$ first implies

$$E[f(W_n(q))] \geq \hat{f}_n \left(\sum_{k=1}^{K} (q^k)^2 \right), \quad q \in \mathcal{S}_K. \tag{5.6}$$

On the other hand, a straighforward Lagrangian argument shows that

$$\min_{q \in \mathcal{S}_K} \sum_{1}^{K} (q^k)^2 = \sum_{1}^{K} \left(\frac{1}{K} \right)^2 = \frac{1}{K} \tag{5.7}$$

and since $\hat{f}_n$ is increasing, we conclude that

$$\hat{f}_n \left(\sum_{k=1}^{K} (q^k)^2 \right) \geq \hat{f}_n \left(\frac{1}{K} \right), \quad q \in \mathcal{S}_K. \tag{5.8}$$

But, upon making use of (5.4)–(5.5), we obtain

$$\hat{f}_n \left(\frac{1}{K} \right) = \sum_{k=1}^{K} \frac{1}{K} \hat{f}_n(\frac{1}{K}) = E[f(W_n(\frac{1}{K} e_K))]. \tag{5.9}$$

Finally, combining (5.6), (5.8) and (5.9), we conclude that

$$E[f(W_n(\frac{1}{K} e_K))] \leq E[f(W_n(q))], \quad q \in \mathcal{S}_K \tag{5.10}$$

for every increasing convex mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$, and (5.2) readily follows. The conclusion (5.3) is now an immediate consequence of (5.2) since here $R_n(q) =_{st} W_n(q) + \sigma_n^1$ for each q in $\mathcal{S}_K$, with $W_n(q)$ and σ_n^1 independent (Ross [10, Prop. 8.5.4, p. 272]).

■

For each $K = 1, 2, \ldots$, we consider a system of K parallel identical queues under the homogeneity assumption (5.1) as described earlier. For each $n = 1, 2, \ldots$, we denote by $W_n^{(K)}$ and $R_n^{(K)}$ the waiting and response times of the nth customer when using the customer assignment characterized by the probability vector $\frac{1}{K} e_K$. In (Makowski and Philips [9]), it was shown that

$$W_n^{(K+1)} \leq_{st} W_n^{(K)}, \quad K = 1, 2, \ldots \qquad n = 0, 1, \ldots \tag{5.11}$$

with a similar result for response times. As a by-product of the proof of Theorem 4, we now obtain the following complement to this monotonicity result.

Theorem 6. *For each $K = 1, 2, \ldots$, assume the conditions* **(A1)–(A4)** *to hold with $q = \frac{1}{K} e_K$. For each $n = 0, 1, \ldots$, $\{W_n^{(K)},\ K = 1, 2, \ldots\}$ and $\{R_n^{(K)},\ K = 1, 2, \ldots\}$ are each SDCX.*

Proof. Fix $n = 0, 1, \ldots$. As in the proof of Theorem 5, we have

$$E[f(W_n^{(K)})] = \hat{f}_n(\frac{1}{K}), \quad K = 1, 2, \ldots \tag{5.12}$$

for every increasing convex mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$. Now, observing that $\frac{1}{K+1} < \frac{1}{K} < \frac{1}{K-1}$ and $\frac{1}{K} - \frac{1}{K+1} < \frac{1}{K-1} - \frac{1}{K}$, we readily obtain the asserted result from the representation (5.12) and the fact that the mapping $\hat{f}_n : \mathbb{R}_+ \to \mathbb{R}_+$ is increasing convex. The proof for the collection $\{R_n^{(K)}, \ K = 1, 2, \ldots\}$ follows similar lines, and is therefore omitted for the sake of brevity.

■

VI. STEADY–STATE RESULTS

In view of Theorem 5, it is natural to wonder whether (5.2)–(5.3) hold true in the sense of the ordering $\leq_{st}$, i.e., for each $n = 0, 1, \ldots$,

$$W_n(\frac{1}{K}e_K) \leq_{st} W_n(q) \quad \text{and} \quad R_n(\frac{1}{K}e_K) \leq_{st} R_n(q), \quad q \in \mathcal{S}_K. \tag{6.1}$$

We were not able to prove (6.1) in general. However, when customers arrive according to a Poisson process, a steady–state version of (6.1) can be obtained as we now show. The key to this result is contained in the following convexity property first announced in (Gün [6,7])

Theorem 7. *Consider a family of $M/GI/1$ queueing systems with arrival rate λ and integrable service time distribution B under the stability condition $\lambda E[B] < 1$. If $W(\lambda)$ and $R(\lambda)$ denote the stationary customer waiting and response times, respectively, then $\{W(\lambda), \ 0 < \lambda < E[B]^{-1}\}$ and $\{R(\lambda), \ 0 < \lambda < E[B]^{-1}\}$ are each SICX(st).*

In (Gün [6,7]), the result was established through direct calculations on phase–type distributions. Here we provide a direct argument based on a probabilistic representation of the stationary waiting in a $M/GI/1$ queue. The discussion uses the following two lemmas which are of independent interest.

Let $\{X_n, \ n = 1, 2, \ldots\}$ be a sequence of i.i.d. $\mathbb{R}_+$–valued rvs. The corresponding partial sums $\{S_n, \ n = 0, 1, \ldots\}$ are defined by $S_n = \sum_{k=1}^{n} X_k$ for all $n = 1, 2, \ldots$, with the usual convention $S_0 = 0$.

Lemma 8. *Let $\{\alpha(\theta), \ \theta \in \Theta\}$ be a collection of $\{0, 1, \ldots\}$–valued rvs, such that for each θ in Θ, the rv $\alpha(\theta)$ is independent of the rvs $\{X_n, \ n = 1, 2, \ldots\}$. If $\{\alpha(\theta), \ \theta \in \Theta\}$ is SICX(st) (resp. SICX), then $\{S_{\alpha(\theta)}, \ \theta \in \Theta\}$ is also SICX(st) (resp. SICX).*

Proof. For any mapping $f : \mathbb{R}_+ \to \mathbb{R}$, we set

$$\tilde{f}(n) := E[f(S_n)], \quad n = 0, 1, \ldots \tag{6.2}$$

The rvs $\{X_n, \ n = 1, 2, \ldots\}$ being non–negative and i.i.d., we see that the mapping $n \to \tilde{f}(n)$ is increasing (resp. integer–convex) if the mapping f is increasing (resp. convex);

details are available in (Ross [10, Lemma 8.6.7, p. 278]). The proof is completed by observing that

$$E[f(S_{\alpha(\theta)})] = E[\tilde{f}(\alpha(\theta))], \quad \theta \in \Theta \tag{6.3}$$

since the rv $\alpha(\theta)$ is assumed independent of the rvs $\{X_n,\ n = 1, 2, \ldots\}$.

■

For each θ in $(0, 1)$, we denote by $\nu(\theta)$ the $\{0, 1, \ldots\}$–valued rv which is geometrically distributed with parameter θ, i.e.,

$$P[\nu(\theta) = n] = (1 - \theta)\theta^n. \qquad n = 0, 1, \ldots \tag{6.4}$$

Lemma 9. *The collection $\{\nu(\theta),\ \theta \in (0, 1)\}$ of geometric rvs defined by (6.4) is SICX(st).*

Proof. For each $n = 0, 1, \ldots$, the mapping $\theta \to P[\theta > n] = (1 - \theta)\theta^{n+1}$ is increasing convex on the interval $(0, 1)$, and the result is now immediate from Proposition 8.1.2 in (Ross [10, p. 252]) and from Lemma 3.3 in (Shaked and Shanthikumar [12, p. 165]).

■

Proof of Theorem 7. It is plain that we need only establish the result for the stationary waiting times. To do so, we shall use a well–known representation for the stationary version of the waiting time in $M/GI/1$ queues. Let $\{X_n,\ n = 0, 1, \ldots\}$ be a sequence of i.i.d. rvs, each one distributed according to the stationary forward recurrence time associated with the the service time distribution B, i.e.,

$$P[X_n \leq x] = \frac{\int_0^x (1 - B(t))dt}{E[B]}, \quad x \geq 0. \qquad n = 0, 1, \ldots \tag{6.5}$$

Now, with the notation of Lemma 9, we can write Eqn. (5.111) of (Kleinrock [8, p. 201]) as

$$W(\lambda) =_{st} S_{\nu(\lambda E[B])}, \quad 0 < \lambda < E[B]^{-1} \tag{6.6}$$

where the geometric rv $\nu(\lambda E[B])$ is independent of the rvs $\{X_n,\ n = 1, 2, \ldots\}$. That $\{W(\lambda),\ 0 < \lambda < E[B]^{-1}\}$ is SICX(st) follows from Lemmas 8 and 9.

■

We now return to the homogeneous system of parallel queues with Bernoulli routing as described in Section 5 under the additional assumption that customers arrive according to a Poisson process with rate λ, i.e., the interarrival time distribution A is given by

$$A(t) = P[\tau_{n+1} \leq t] = 1 - e^{-\lambda t}, \quad t \geq 0. \qquad n = 0, 1, \ldots \tag{6.7}$$

A stationary version of (6.1) is now presented in Theorem 10 where we use the notation

$$\mathcal{S}_K^* := \{q \in \mathcal{S}_K : \lambda q^k E[B] < 1,\ 1 \leq k \leq K\}. \tag{6.8}$$

Theorem 10. *Assume* **(A1)–(A4)** *with the homogeneity assumption (5.1) and the Poisson assumption (6.7). If* $\lambda E[B] < K$, *then the stochastic ordering relations*

$$W_\infty(\frac{1}{K}e_K) \leq_{st} W_\infty(q), \quad q \in \mathcal{S}_K^* \tag{6.9}$$

and

$$R_\infty(\frac{1}{K}e_K) \leq_{st} R_\infty(q), \quad q \in \mathcal{S}_K^* \tag{6.10}$$

hold true, i.e., the routing assignment $q^* = \frac{1}{K}e_K$ *minimizes the stationary waiting and response times in the sense of the ordering* $\leq_{st}$.

Note that the condition $\lambda E[B] < K$ ensures that the probability vector q^* lies in the set $\mathcal{S}_K^*$. Moreover, only probability vectors in $\mathcal{S}_K^*$ need to be considered since for q in $\mathcal{S}_K$ but not in $\mathcal{S}_K^*$, at least one of the K parallel queues is unstable and (6.9)–(6.10) automatically holds.

Proof of Theorem 10. Fix $\lambda > 0$ and q in $\mathcal{S}_K^*$, in which case each one of the K parallel queues is stable and the various quantities of interest admit stationary versions. In particular, with the notation of Theorem 7, we have $V_\infty^k(q^k) =_{st} W(\lambda q^k)$, $1 \leq k \leq K$ (Makowski and Philips [9]).

As in the proof of Theorem 4, in analogy with (4.9), for every Borel mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$, we have

$$\begin{aligned} E[f(W_\infty(q))] &= \sum_{k=1}^K q^k E[f(V_\infty^k(q^k))] \\ &= \sum_{k=1}^K q^k E[f(W(\lambda q^k))]. \end{aligned} \tag{6.11}$$

If f is increasing, then the mapping $(0, \min\{1, (\lambda E[B])^{-1}\}) \to \mathbb{R}_+ : q^k \to E[f(W(\lambda q^k)]$ is increasing convex by virtue of Theorem 7. The arguments of Theorem 5 now imply that

$$E[f(W_\infty(\frac{1}{K}e_K))] \leq E[f(W_\infty(q))] \tag{6.12}$$

and the result (6.9) follows. The proof of (6.11) is conducted similar lines and is left to the interested reader.

■

VII. ADDITIONAL COMMENTS

In this final section we have collected several comments on the results discussed earlier in this paper.

1. A careful inspection of the inductive proof of Theorem 1 reveals that Theorem 4 will hold true under conditions weaker than the assumed conditions **(A1)–(A4)**. In fact,

the condition **(A2)** on the arrival sequence $\{\tau_{n+1},\ n = 0,1,\ldots\}$ can be safely dropped, and conditions **(A1)**, **(A3)** and **(A4)** alone (with no further assumption on the arrival stream) will ensure the validity of Theorem 4. This can be seen by making the following modification in the induction argument in Theorem 1 when applied to the Lindley recursion (4.3); clearly only Claim 1 of Theorem 4 needs to be established: Fix $1 \leq k \leq K$. If for some $n = 0,1,\ldots$, $\{V_n^k(q^k),\ q^k \in [0,1]\}$ is SICX(sp), so is $\{V_n^k(q^k) - \tau_{n+1},\ q^k \in [0,1]\}$ by the independence assumption **(A1)** (Shaked and Shanthikumar [11, Thm. 3.10, p. 436]), while $\{u_n^k(q^k)\sigma_n^k,\ q^k \in [0,1]\}$ was argued earlier to be SICX(sp). Using again independence, we can conclude that $\{V_n^k(q^k) + u_n^k(q^k)\sigma_n^k - \tau_{n+1},\ q^k \in [0,1]\}$ is SICX(sp), and so is $\{V_{n+1}^k(q^k),\ q^k \in [0,1]\}$ in view of (4.3) (Shaked and Shanthikumar [11, Thm. 3.2, p. 433]). Theorems 5 and 6 also hold under the weaker set of conditions **(A1)**, **(A3)** and **(A4)**, since these propositions are direct consequences of Theorem 4.

2. The methods of this paper can be used to obtain versions of Theorems 4–6 for more complex situations. For instance, each one of the K parallel $GI/GI/1$ queues can be replaced by a tandem series of single server queues, the kth one being composed of I_k service stations in series. In that case the Lindley recursions (4.3) needs to be replaced by a system of recursions which we now describe: For $n = 0,1,\ldots$, let $D_n^{k,i}(q)$ denote the service completion time of the nth customer at the ith service station if it had been routed to the kth channel (or tandem series), $1 \leq k \leq K$ and $1 \leq i \leq I_k$. Setting

$$D_n^{k,0}(q) = 0, \quad D_{n+1}^{k,0}(q) = D_n^{k,0}(q) + \tau_{n+1}, \quad 1 \leq k \leq K \qquad n = 0,1,\ldots (7.1)$$

we find that

$$\begin{aligned} D_{n+1}^{k,i+1}(q) &= \max\{D_{n+1}^{k,i}(q), D_n^{k,i+1}(q)\} + u_n^k(q)\sigma_n^{k,i+1} \\ D_0^{k,i+1}(q) &= 0, \quad 1 \leq k \leq K,\ 1 \leq i \leq I_k \end{aligned} \qquad n = 0,1,\ldots (7.2)$$

where $\sigma_n^{k,i}$ denotes the duration of the service requirement provided to the nth customer at the ith service station if it had been routed to the kth channel. The customer response time $R_n(q)$ of the nth customer is simply given by

$$R_n(q) = \sum_{k=1}^{K} u_n^k(q) \cdot \left(D_n^{k,I_k}(q) - D_n^{k,0}(q)\right). \qquad n = 0,1,\ldots (7.3)$$

3. Steady–state versions of Theorems 4–6 are easily obtained whenever applicable. In that respect, Theorem 10 is a strenghtening of a steady–state version of Theorem 5 in that the comparison holds in the stronger ordering $\leq_{st}$, and not merely in the ordering $\leq_{icx}$.

4. As in Theorem 7, consider a family of $M/GI/1$ queues with arrival rate λ and service distribution B. For $t \geq 0$, let $W_t(\lambda)$ denote the virtual waiting time (or workload) at time t in the $M/GI/1$ queue with arrival rate λ. It is known (Chang, Cao and Pinedo [2, Lemma 3.2]) that for each $t \geq 0$, the collection of virtual times $\{W_t(\lambda),\ \lambda \geq 0\}$ is SICX(sp). Note however that Theorem 7 is not a consequence of this result, despite the well–known fact that under the stability condition $\lambda E[B] < 1$, we have $W_t(\lambda) \Longrightarrow W(\lambda)$ (as t goes to ∞) where as before $W(\lambda)$ denotes the stationary customer waiting time. Indeed, it is not known wether SICX(sp) is stable under weak convergence. Moreover, even if it were the case, the stronger property SICX(st) is not implied by SICX(sp).

REFERENCES

[1] 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.

[2] C.S. Chang, X.L. Chao and M. Pinedo, "Monotonicity results for queues with doubly stochastic Poisson arrivals: Ross's conjecture," *Journal of Applied Probability*, forthcoming (1990).

[3] C.S. Chang, X.L. Chao and M. Pinedo, "A note on queues with Bernoulli routing," IBM Research Report **RC 15418** (1990).

[4] C.S. Chang, "A new ordering for stochastic majorization: Theory and applications," Preprint (1990).

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

[6] L. Gün, *Performance Evaluation and Optimization of Parallel Systems with Synchronization,* Ph.D. Thesis, Electrical Engineering Department, University of Maryland, College Park (MD), August 1989. Available as SRC Technical Report **89–64**, Systems Research Center, University of Maryland, College Park (MD) (1989).

[7] L. Gün, "On the waiting and response time distributions of the $M/GI/1$ queue," in *Proceedings of BILCON '90*, Ankara (Turkey), July 1990.

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

[9] A.M. Makowski and T.K. Philips, "Simple proofs of some folk theorems for parallel queues," IBM Research Report **RC 14556** (1989).

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

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

[12] M. Shaked and J.G. Shanthikumar, "Convexity of a set of stochastically ordered random variables," *Advances in Applied Probability* **22** (1990), pp. 160–177.

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

[14] P. Tsoucas, "Stochastic monotonicity of the output process of parallel queues," Preprint (1989).