The fork-join queue and related systems with synchronization constraints: stochastic ordering,approximations and computable bounds

François Baccelli, A. Makowski, A. Shwartz

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tel.(1) 39 63 55 11

Rapports de Recherche

N° 687

THE FORK-JOIN QUEUE AND RELATED SYSTEMS WITH SYNCHRONIZATION CONSTRAINTS: STOCHASTIC ORDERING, APPROXIMATIONS AND COMPUTABLE BOUNDS

François BACCELLI
Armand M. MAKOWSKI
Adam SHWARTZ

JUIN 1987

DRAFT - January 1987

THE FORK-JOIN QUEUE AND RELATED SYSTEMS WITH SYNCHRONIZATION CONSTRAINTS: STOCHASTIC ORDERING, APPROXIMATIONS AND COMPUTABLE BOUNDS

by

Francois Baccelli[1], Armand M. Makowski[2] and Adam Shwartz[3]

ABSTRACT

A simple queueing system, known as the Fork-Join queue, is considered with basic performance measure defined as the delay between the Fork and Join dates. Simple lower and upper bounds are derived for some of the statistics of this quantity. They are obtained, in both transient and steady-state regimes, by stochastically comparing the original system to other queueing systems with a structure simpler than the original system, yet with identical stability characteristics. In steady-state, under renewal assumptions, the computation reduces to standard GI/GI/1 calculations and the approximations constitute a first sizing-up of system performance. These bounds can also be used to show that for homogeneous Fork-Join queue systems under renewal assumptions, the moments of the system response time grow logarithmically in the number of parallel processors. The bounding arguments make use of ideas from the theory of stochastic ordering and of the notion of associated random variables. They are of independent interest to study various other queueing systems with synchronization constraints.

---

[1] INRIA, Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay, Cedex, France. The work of this author was supported partially through a grant from the Minta Martin Aeronautical Research Fund, College of Engineering, University of Maryland at College Park.

[2] Electrical Engineering Department, University of Maryland, College Park, Maryland 20742, U.S.A. The work of this author was supported partially through ONR Grant N00014-84-K-0614, partially through NSF Grant ECS-83-51836 and partially through a grant AT&T Bell Laboratories.

[3] Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa 32000, Israel. The work of this author was supported through a Grant from AT&T Bell Laboratories.

FILES D'ATTENTE AVEC DES SYNCHRONISATIONS DE TYPE FORK-JOIN : ORDONNANCEMENT STOCHASTIQUE, APPROXIMATIONS ET BORNES

François BACCELLI[1], Armand M. MAKOWSKI[2], Adam SHWARTZ[3]

Résumé

Cet article présente une étude des statistiques des temps de réponse dans un réseau de files d'attente en parallèle synchronisées par un mécanisme de type "Fork-Join". Des théorèmes d'ordonnancement stochastique sont utilisés pour établir des bornes supérieures et inférieures à forme produit sur les moments de ces temps de réponse. En régime stationnaire et sous des hypothèses de renouvellement, le calcul de ces bornes se réduit au calcul de simples files GI/GI/I. Dans le cas homogène, on déduit aussi de ces bornes des asymptotiques logarithmiques pour les moments lorsque la taille du réseau devient grande. Les méthodes employées se fondent sur l'ordonnancement convexe et sur la notion de variables aléatoires associées et sont applicables à d'autres systèmes de files d'attente synchronisées.

[1] INRIA, Domaine de Voluceau, Rocquencourt, B.P.105, 78153 Le Chesnay Cedex (France).

[2] Electrical Engineering Department, University of Maryland, College Park, Maryland 20742 (U.S.A).

[3] Electrical Engineering Department, Technion-Israel Institute of Technology, Haifa 32000 (Israel).

I. INTRODUCTION

A K-dimensional Fork-Join (FJ) queue is a queueing system operated by parallel K servers with *synchronized* arrival streams. Each server is attended by a buffer of infinite capacity and individually operates according to the FIFO discipline. Customers arrive into the system in bulks of size not larger than K and are processed according to the following discipline.

Upon arrival, a bulk of size $S \leq K$, bringing customers to S of the K servers, is immediately split so that each one of the S customers composing it is allocated to its server (the so-called Fork Primitive).

As soon as all the S customers constituting a bulk have been serviced, the bulk is immediately recomposed (the so-called Join primitive) and leaves the system at once. This second synchronization constraint is achieved by parking already serviced customers in an auxiliary buffer of infinite capacity, where they await being reunited to serviced customers of the same bulk whose service has not been completed yet.

Such queueing models arise in many application areas, including flexible manufacturing and parallel processing (e.g. the cobegin and coend structures in concurrent languages), with a wide variety of interpretations. In the context of production systems, a bulk customer can be interpreted as a customer's order with several components, each component or suborder being attended by a separate production device. An example very similar to this one is obtained by considering the production of multipart items. In computer systems with parallel architecture, a bulk customer can be viewed as a program composed of several subroutines, each one to be executed on a different processor.

For this type of applications, the determination of bulk response time (defined as the delay between the Fork and the Join dates) is of crucial importance in quantifying system performance. In two dimensions ($K = 2$) and in the case of constant bulk sizes $S = K = 2$, the stationary joint distribution of the number of customers in the two queues was determined by Flatto and Hahn [12] under Markovian assumptions. A somewhat more general problem with Poisson arrivals and general service times was also analyzed by Baccelli in [1]. However, in more dimensions (i. e., $K > 2$) and/or for more general interarrival distributions, the problem still remains open.

In this paper, simple lower and upper bounds are derived for various statistics of this response time, including its moments. In order to carry out this program, it is convenient (and natural) to consider a somewhat larger class of queueing systems with K parallel servers with synchronization constraints. This enlarged class is obtained by removing the synchronization constraint on the

arrival streams and by allowing more general loading patterns. This class of models includes the simplest FJ queue studied earlier by the authors [2,5], and is described in detail in Section 2 where the performance measures of interest are defined.

The bounds derived in this paper have both transient and steady-state versions. They are obtained by a direct stochastic comparison of the queueing system to several other systems with K parallel servers, which exhibit stability conditions identical to the one for the original system.

Two basic bounding methodologies are used throughout this work. The first approach relies on the *convex increasing* order for probability distributions [19,21] and is found most useful for establishing monotonicity results on the sequence of waiting times which are generated through Lindley's equation. A basic bounding result is obtained for such a recursion in Section 3, and the ideas are then applied in Section 4 to the queueing system of interest. Preliminary versions of the results derived here can be found in the conference papers [2,5] for the FJ queue under renewal type assumptions.

The second methodology makes use of the notion of *associated* random variables [8,11] and is especially useful in establishing comparison results for the *maximum* of correlated random variables. The necessary material is developped in Section 5, where it is shown roughly speaking that increasing the number of independent components increases the system response time. This technique was already used by Nelson and Tantawi [17] who obtained only first moment information in the special case of Poisson arrivals and exponential servers.

In Section 6, the bounds are then specialized to the FJ queue with synchronized arrivals under a set of renewal assumptions. The discussion emphasizes the *computability* of the steady-state bounds, in the sense that their evaluation reduces to analyzing the statistics of K *independent* GI/GI/1 queueing systems, in contrast with the initial problem which is also K dimensional but with *strongly coupled* components. The calculations are carried out explicitly for the case of *homogeneous* FJ queues, i. e., the loading patterns and service requirements are identical for all processors, when the service times are *exponentially* distributed.

A further use of these bounds is discussed in Section 7 where homogeneous FJ queues are considered. Asymptotics on the system response time statistics are obtained as the number K of servers grows large. It is shown under standard renewal assumptions that the moments of the system response time grow *logarithmically* in the number of parallel servers, provided the Laplace-Stieltjes transform of the service time distribution is *rational*. This result, announced in [3], generalizes a similar result of Nelson and Tantawi [17] for the special case of Poisson arrivals and exponential

servers. The asymptotics are developped by making use of a not too well-known result of Lai and Robbins [15] on the maximum of a class of dependent random variables.

It should be noted that better bounds or more precise asymptotics could be developed for special situations by making use of the ideas discussed in this paper. The choice of not to do so was guided by the concern of not lengthening further an already lengthy manuscript. However, the bounding methodologies discussed here are of independent interest in the analysis of other queueing systems with synchronization constraints. The first one, based on the convex increasing ordering, was applied successfully by Baccelli and Massey in the study of parrallel and/or series networks of FJ queues [7].

2. THE MODEL

The queueing model of interest in this paper is presented in this section, together with the notation and some of the basic assumptions enforced throughout. The queueing model which is now introduced is far more general than synchronized FJ queue model that motivated this work [2,5]; its usefulness will become apparent to the reader in the discussion carried out in the next sections.

The basic random variables

Emphasis is put on sample path representation for the quantities of interest and as further developments will demonstrate, this approach is quite fruitful in establishing bounds. To that end, an underlying probability triple $(\Omega, \mathcal{F}, P)$ is postulated on which all the random variables (RV) mentioned in this paper are defined. A positive integer K is given and held fixed hereafter. As a convention, the k-th component RV of any $I\!R^K$-valued RV is denoted by the same symbol as this RV but superscripted by k; a similar convention is adopted to denote the components of any vector in $I\!R^K$. This probability triple $(\Omega, \mathcal{F}, P)$ is assumed to simultaneously carry the sequences $\{\tau_n\}_1^\infty$, $\{\sigma_n\}_0^\infty$ and $\{u_n\}_0^\infty$ of $I\!R_+^K$-valued RV's, together with an $I\!R_+^K$-valued RV W.

The queueing system generated by the constituting sequence

$$\langle W, \sigma_n, u_n, \tau_{n+1},\ n = 0, 1, ... \rangle \tag{2.1}$$

is defined as a queueing system composed of K ***parallel*** servers with the following features: Each one of these servers has its own waiting area of ***infinite*** capacity and operates according to the FIFO discipline. The RV's $\{\tau_n^k\}_1^\infty$ model the interarrival times as experienced by the k-th server, with arrivals to the corresponding queue thus taking place along the time sequence $\{A_n^k\}_0^\infty$ defined

by

$$A_n^k = \sum_{m=0}^{n-1} \tau_{m+1}^k \qquad n = 1, 2...(2.2)$$

with $A_0^k = 0$. The customers arriving at the k-th queue at such times are called type k customers hereafter. The n-th customer of type k brings an amount of processing time σ_n^k to be executed by the k-th server, whereas the RV u_n^k represents a *loading* factor in that the actual work to be processed by the k-th server due to the n-th arrival of type k is $u_n^k \cdot \sigma_n^k$. A customer is assumed to arrive at time $t = 0$, at which time an initial load is already awaiting service in the various buffer areas, and the RV W^k thus represents the amount of time required by the k-th server to clear this initial load from the k-th queueing area.

The following special cases will be of special interest in the sequel:

(C.1): The RV's $\{u_n\}_0^\infty$ take their values in $\{0,1\}^K$. In that case, each RV u_n determines a (random) subset I_n of $\{1, 2..., K\}$ given by

$$I_n := \{k, 1 \leq k \leq K : u_n^k = 1\}, \qquad n = 0, 1, ...(2.3)$$

and the system is one fed by K arrival streams, where the n-th (composite) arrival brings work only to the subset I_n of the K servers, possibly not all at the same time.

(C.2): The RV's $\{u_n\}_0^\infty$ are *synchronized* in the sense that

$$u_n^1 = u_n^2 = ... = u_n^K = v_n. \qquad n = 0.1, ...(2.4)$$

In this system, the amounts of actual work brought by the n-th customers to the K parallel queues are positively correlated through a common scaling factor v_n. Such a correlation seems quite natural in all the practical interpretations discussed earlier (e.g., the sizes of parallel subprograms in a program).

2.2 The performance measures

In order to define reasonable performance measures, consider the sequence of $\mathbb{R}_+^K$-valued RV's $\{W_n\}_0^\infty$ generated componentwise by the recursions

$$W_{n+1}^k = [W_n^k + u_n^k \cdot \sigma_n^k - \tau_{n+1}^k]^+, \quad 1 \leq k \leq K, \qquad n = 0, 1, ...(2.5)$$

with $W_0 = W$. The RV W_n^k represents the *waiting time* of the n-th customer of type k, whereas the quantities R_n^k and S_n^k, akin to response times in the queueing system attended by the k-th server,

are defined by

$$R_n^k := u_n^k \cdot (W_n^k + \sigma_n^k), \quad 1 \le k \le K, \qquad n = 0, 1, ...(2.6)$$

and

$$S_n^k := W_n^k + u_n^k \cdot \sigma_n^k, \quad 1 \le k \le K, \qquad n = 0, 1, ...(2.7)$$

respectively. For reasons that will become clear later on, two quantities can be naturally defined as the *system response time* for the n-th (composite) customer; these are denoted $\{T_n^1\}$ and $\{T_n^2\}$, and are given by

$$T_n^1 := \max_{1 \le k \le K} R_n^k = \max_{1 \le k \le K} u_n^k \cdot (W_n^k + \sigma_n^k) \qquad n = 0, 1, ...(2.8)$$

and

$$T_n^2 := \max_{1 \le k \le K} S_n^k = \max_{1 \le k \le K} W_n^k + u_n^k \cdot \sigma_n^k, \qquad n = 0, 1, ...(2.9)$$

respectively.

The definitions (2.8)-(2.9) attempt to provide meaningful measures of system performance for as large a class of models as possible. To get a better feel as to the meaning of these definitions, consider the situation where the arrivals are *synchronized* in the sense that

$$\tau_n^1 = \tau_n^2 = ... = \tau_n^K \qquad n = 0, 1, ...(2.10)$$

This corresponds to the situation where the n-th customers of all types arrive into the system at the same time, in which case the arrival stream *common* all K queues is still denoted by $\{\tau_{n+1}\}_0^\infty$. In the particular case (C.1), T_n^1 now reads

$$T_n^1 = \max_{k \in I_n} R_n^k \qquad n = 0, 1, ...(2.11)$$

with I_n given by (2.2), and is exactly the time that elapses between the Fork and the Join dates of the n-th arrival. On the other hand, when the loading sequence has the form (2.4) as in (C.2), it is more natural to define the system response time of the n-th customer by (2.9).

The system response times defined through (2.8) and (2.9) coincide when the loading sequence has the simplified form

$$u_1^k = \cdots = u_n^K = 1 \qquad n = 0, 1, ...(2.12)$$

as in the simplest FJ queue system studied earlier by the authors [2,5]. The difficulty of analyzing these queueing systems with synchronization constraints is already apparent in the simple situation

defined by (2.10) and (2.12), say under standard renewal assumptions [2,5]. Indeed, in such a case, the single server queueing system associated with each server embedded in the FJ queue operates like a standard GI/GI/1 queueing system. However, these K parallel GI/GI/1 systems are *not* independent in general since they have *identical* inputs owing to (2.10). It is this very lack of independence that makes the computation of the statistics of the RV's $\{T_n^i\}_0^\infty$, $i = 1,2$, extremely hard. In view of these difficulties, it seems relevant to seek ways of generating *bounds* and *approximations* to the various statistics of $\{T_n^i\}_0^\infty$, $i = 1,2$. The results on stochastic ordering presented in this paper readily lead to the derivation of *simple computable* bounds to these statistics in the context of the FJ queue as well as for the more general queueing systems described earlier in this section. The *comparison* results discussed in this paper are derived through direct bounding arguments that explicitly exploit the sample path nature of the recursions (2.5). The basic idea consists in *directly* constructing from the RV's that define the original system, a new queueing system of the same type but with different constituting sequence (2.1)

A few remarks are now in order concerning the situation where (C.1) holds. With the notation (2.3) it is convenient to say that the n-th bulk (or composite customer) is of type I if $I_n = I$, where I is a subset of $\{1,\cdots,K\}$. In the synchronized case (2.10), there is a single stream which includes bulk arrivals of all possible types. Another model would consist in defining one arrival process for each bulk type. Presently, this modification in the model seems immaterial. However, all further developments leading to the derivation of bounds will require that the interarrival times sequence $\{\tau_{n+1}\}_0^\infty$ be independent of the type sequence $\{I_n\}_0^\infty$. Even under standard renewal assumptions, this restriction introduces a clear difference between these two models of arrival patterns. The generalization of the bounds derived in the paper to such multiclass arrival patterns is still an open problem.

2.3 Notation

Although more general situations will be covered during the discussion, the results will often be specialized to various models of interest. In particular, it will be convenient to refer to the system under the (resp. *strong*) *independence* assumption (I) (resp. (Ibis)) if the conditions below are satisfied, namely

(I) All the IR_+^K-valued RV's $\{W, \sigma_n, u_n, \tau_{n+1},\ n = 0,1,...\}$ are ***mutually independent***; and

(Ibis) All the IR_+-valued RV's $\{W^k, \sigma_n^k, u_n^k, \tau_{n+1}^k,\ 1 \le k \le K,\ n = 0,1,...\}$ are ***mutually independent.***

This section closes with a word on the notation used throughout the paper: The conditional

expectation of any IR-valued RV X with respect to any sub-σ-field $\mathcal{D}$ of $\mathcal{F}$ is often denoted by $X^{\mathcal{D}}$, with

$$X^{\mathcal{D}} := E[X \mid \mathcal{D}], \tag{2.13}$$

whenever meaningful. The notations $\sigma_n^{k,\mathcal{D}}$, $W_n^{k,\mathcal{D}}$, $R_n^{k,\mathcal{D}}$, $S_n^{k,\mathcal{D}}$ and $T_n^{i,\mathcal{D}}$ are thus used with this meaning for all $n = 0, 1, ...,\ 1 \leq k \leq K$, and $i = 1, 2$.

If the RV X has probability distribution function $F(\cdot)$, then its Laplace-Stieltjes transform $F^*(\cdot)$ is given by

$$F(s) = E[e^{-sX}] = \int_{IR} e^{-sx} dF(x) \tag{2.14}$$

for s in some appropriate region of convegence.

3. BOUNDS IN THE CONVEX INCREASING ORDERING

The first bounding methodology is based on a simple result that provides for a *direct stochastic* comparison between different queueing systems of the type described earlier. This elementary result uses in an essential way the structure of the Lindley's recursions (2.5), and allows for a *unified* treatment of many of the bounds presented here. The discussion that follows finds its origin in a folk theorem of Queueing Theory stating that *determinism minimizes waiting (resp. response) times* in many queueing systems. For G/G/1 systems, such results have been established under a variety of assumptions by a number of authors, including Hajek [13], Humblet [14] and Rogozin [18] to name a few.

3.1 A basic bounding methodology

To set the stage for the discussion, consider a sequence $\{O_n\}_1^\infty$ of IR-valued RV's and an IR_+-valued RV V defined on the probability triple $(\Omega, \mathcal{F}, P)$. These RV's are assumed to satisfy the *finite* mean condition

$$E[V] < \infty, \quad E[O_n] < \infty. \qquad n = 1, 2, ... \tag{3.1}$$

For any σ-field $\mathcal{D}$ of events contained in $\mathcal{F}$, the sequence $\{O_n^{\mathcal{D}}\}_1^\infty$ of IR-valued RV's is defined to be (any one version of) the conditional expectations

$$O_n^{\mathcal{D}} := E[O_n \mid \mathcal{D}], \qquad n = 1, 2, ... \tag{3.2}$$

in agreement with the convention (2.13). The main result of this section is then contained in

Theorem 3.1. *Let $\{V_n\}_0^\infty$ and $\{V_n(\mathcal{D})\}_0^\infty$ be the sequences of IR_+-valued RV's defined through the recursions*

$$V_{n+1} = [V_n + O_{n+1}]^+ \qquad n = 0, 1, ... \tag{3.3}$$

and

$$V_{n+1}(\mathcal{D}) = [V_n(\mathcal{D}) + O_{n+1}^{\mathcal{D}}]^+ \qquad n = 0, 1, ...(3.4)$$

with $V_0 = V_0(\mathcal{D}) = V$. *Under the assumptions made, whenever the RV* V *is* $\mathcal{D}$*-measurable, the inequalities*

$$V_n(\mathcal{D}) \leq E[V_n \mid \mathcal{D}] \quad a.s \qquad n = 0, 1, ...(3.5)$$

hold.

Proof: The proof proceeds by induction. Since the RV V is $\mathcal{D}$-measurable, (3.5) trivially holds for $n = 0$ since the initial term is V in both recursions.

Take as induction hypothesis that (3.5) holds true for *some* $n = m$. For such m, Jensen's inequality gives

$$E[V_{m+1} \mid \mathcal{D}] \geq [E[V_m \mid \mathcal{D}] + E[O_{m+1} \mid \mathcal{D}]]^+ \qquad (3.6)$$

since the function $IR \rightarrow IR : x \rightarrow x^+$. is convex monotone non-decreasing. Substitution of (3.2) into (3.6) and use of the induction hypothesis then yield the a.s relations

$$E[V_{m+1} \mid \mathcal{D}] \geq [V_m(\mathcal{D}) + O_{m+1}^{\mathcal{D}}]^+ = V_{m+1}(\mathcal{D}), \qquad (3.7)$$

where the last equality follows from (3.4). This shows that (3.5) holds for $n = m + 1$ and since it holds for $n = 0$, it holds by induction for all $n = 0, 1, ...$ □

The following corollary is an easy consequence of Theorem 3.1.

Corollary 3.2 *If* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *are two* σ*-fields of events such that*

$$\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \mathcal{F}, \qquad (3.8)$$

then, with the notation of Theorem 3.1, the a.s. inequalities

$$V_n(\mathcal{D}_1) \leq E[V_n(\mathcal{D}_2) \mid \mathcal{D}_1] \leq E[V_n \mid \mathcal{D}_1] \qquad n = 0, 1, ...(3.9)$$

hold provided $V_0 = V_0(\mathcal{D}_1) = V_0(\mathcal{D}_2) = V$ *and* V *is* $\mathcal{D}_1$*-measurable.*

Proof: From the definition (3.2), the smoothing property for conditional expectations implies that

$$E[O_n^{\mathcal{D}_2} \mid \mathcal{D}_1] = O_n^{\mathcal{D}_1} \quad a.s. \qquad n = 1, 2, ...(3.10)$$

owing to the inclusion (3.8). The first inequality in (3.9) is now a straightforward consequence of Theorem 3.1 when applied to the recursion (3.3)-(3.4) with driving sequences $\{0_n^{\mathcal{D}_1}\}_1^\infty$ and $\{0_n^{\mathcal{D}_2}\}_1^\infty$,

respectively. The second inequality in (3.9) follows from (3.5) (with $\mathcal{D}_2$ replacing $\mathcal{D}$) upon conditioning with respect to the σ-field $\mathcal{D}_1$. □

3.2 Transient Analysis

The basic results of the previous section are now applied to the K server queueing system generated by the constituting sequence

$$\langle W, \sigma_n, u_n, \tau_{n+1},\ n = 0, 1, ... \rangle \tag{3.11}$$

as described in Section 2. The discussion is given under the following two assumptions (A.1)-(A.2), where

(A.1) For all $1 \leq k \leq K$, the RV's $\{\sigma_n^k\}_0^\infty$, $\{u_n^k\}_0^\infty$ and $\{\tau_{n+1}^k\}_0^\infty$ have *finite* means.

(A.2) There exists a sub σ-field $\mathcal{D}$ of $\mathcal{F}$ with the property that W is $\mathcal{D}$-measurable and for each $n = 0, 1, ...$, the RV u_n is *conditionally independent* of the σ-field $\mathcal{G}_n$ *given* $\mathcal{D}$, with

$$\mathcal{G}_n := \sigma\{W, \sigma_n\} \vee \sigma\{\sigma_m, u_m, \tau_{m+1}, 0 \leq m < n\} \qquad n = 0, 1, ...$$

Note that (A.2) is automatically satisfied for any σ-field $\mathcal{D}$ when the consituting loading sequence has the simple form (2.12) as in [2,5] or when the independence condition (Ibis) holds.

With a notation consistent with that introduced in Theorem 3.1, for any σ-field $\mathcal{D}$, let $\{W_n(\mathcal{D})\}_0^\infty$, $\{R_n$ and $\{T_n^i(\mathcal{D})\}_0^\infty$ be the RV's defined through (2.5)-(2.9), respectively, for the K server queueing system generated by the constituting sequence

$$\langle W, \sigma_n^{\mathcal{D}}, u_n^{\mathcal{D}}, \tau_{n+1}^{\mathcal{D}},\ n = 0, 1, ... \rangle. \tag{3.12}$$

Theorem 3.3 *Under the assumptions* (A1)-(A2), *for all* $1 \leq k \leq K$, *the a.s. inequalities*

$$W_n^k(\mathcal{D}) \leq W_n^{k,\mathcal{D}} \qquad n = 0, 1, ... \tag{3.13}$$

$$R_n^k(\mathcal{D}) \leq R_n^{k,\mathcal{D}} \qquad n = 0, 1, ... \tag{3.14}$$

and

$$S_n^k(\mathcal{D}) \leq S_n^{k,\mathcal{D}} \qquad n = 0, 1, ... \tag{3.15}$$

hold, whence

$$T_n^i(\mathcal{D}) \leq T_n^{i,\mathcal{D}}, \quad i = 1, 2. \qquad n = 0, 1, ... \tag{3.16}$$

Proof. For all $n = 0, 1, ...$, the RV W_n is $\mathcal{G}_n$-measurable, and by assumption (A.2), the RV u_n is thus conditionally independent of the RV's $\{\sigma_n, W_n\}$ given the σ-field $\mathcal{D}$, whence

$$E[u_n^k \cdot \sigma_n^k \mid \mathcal{D}] = u_n^{k,\mathcal{D}} \cdot \sigma_n^{k,\mathcal{D}} \quad a.s. \qquad n = 0, 1, ...(3.17a)$$

and

$$E[u_n^k \cdot W_n^k \mid \mathcal{D}] = u_n^{k,\mathcal{D}} \cdot W_n^{k,\mathcal{D}} \quad a.s. \qquad n = 0, 1, ...(3.17b)$$

for all $1 \leq k \leq K$.

With the IR_+^K-valued RV's $\{O_{n+1}\}_0^\infty$ defined componentwise by

$$O_{n+1}^k := u_n^k \cdot \sigma_n^k - \tau_{n+1}^k \qquad n = 0, 1, ...(3.18)$$

for all $1 \leq k \leq K$, it is plain from (3.17a) that

$$O_{n+1}^{k,\mathcal{D}} = u_n^{k,\mathcal{D}} \cdot \sigma_n^{k,\mathcal{D}} - \tau_{n+1}^{k,\mathcal{D}} \qquad n = 0, 1, ...(3.19)$$

for all $1 \leq k \leq K$. The inequality (3.13) now follows readily from Theorem 3.1 applied to the recursions (2.5)-(2.9), whereas (3.14) and (3.15) are direct consequences of (3.13) and (3.17). It is also plain that

$$E[T_n^1 \mid \mathcal{D}] \geq \max_{1 \leq k \leq K} E[R_n^k \mid \mathcal{D}] \geq \max_{1 \leq k \leq K} R_n^k(\mathcal{D}) \quad a.s. \qquad n = 0, 1, ...(3.20)$$

where (3.14) was used in the last inequality, and the validity of (3.16) for the case $i = 1$ is now established. The proof for the case $i = 2$ is similar and is therefore omitted. □

A careful inspection of the proof of Theorem 3.3 reveals that only the consequences (3.17) of the assumption (A.2) enforced on the σ-field $\mathcal{D}$ are needed to carry out the arguments. In fact, (3.17a) alone implies (3.13) and (3.15). The following corollary is an easy by-product of Theorem 3.3.

Corollary 3.4. *Under the assumptions* (A.1)-(A.2), *the inequalities*

$$E[T_n^i(\mathcal{D})] \leq E[T_n^i], \quad i = 1, 2 \qquad n = 0, 1, ...(3.21)$$

hold; more generally, for all convex monotone non-decreasing function $\phi : IR \to IR$,

$$E[\phi(T_n^i(\mathcal{D}))] \leq E[\phi(T_n^i)], \quad i = 1, 2 \qquad n = 0, 1, ...(3.22)$$

provided both expectations exist.

Proof. It suffices to establish (3.22) as (3.21) follows from it (by taking $\phi(x) = x$). Theorem 3.3 yields

$$T_n^i(\mathcal{D}) \leq E[T_n^i \mid \mathcal{D}], \quad i = 1,2 \qquad n = 0,1,...(3.23)$$

and Jensen's inequality thus implies the inequalities

$$\phi(T_n^i(\mathcal{D})) \leq E[\phi(T_n^i) \mid \mathcal{D}], \quad i = 1,2 \qquad n = 0,1,...(3.24)$$

provided these expectations are well defined. The conclusion (3.22) is now immediate from (3.24) upon taking the mathematical expectation on both sides of these inequalities. □

The comparison results of Theorem 3.3 and of its Corollary 3.4 are really statements on the *convex increasing ordering* between waiting times and response times in two different queueing systems, as understood by Stoyan [21], Whitt [22] and many other authors. More precisely, let X and Y be any two IR^K-valued RV's defined on Ω. The (distribution of the) RV X is said to be *greater* than the (distribution of the) RV Y in the *(stochastic) convex increasing order* if and only if

$$E[\phi(Y)] \leq E[\phi(X)] \qquad (3.25)$$

for all *convex monotone non-increasing* mapping $\phi : IR^K \to IR$ for which (3.25) makes sense; this is denoted in short by $X \leq_{ci} Y$.

With this notation, Corollary 3.3 can be restated simply as saying that

$$T_n^i(\mathcal{D}) \leq_{ci} T_n^i, \quad i = 1,2. \qquad n = 0,1,...(3.26)$$

In fact, through an argument identical to the one made in Corollary 3.4, the inequalities (3.13)-(3.14) imply the following stronger stochastic ordering in *vector* form.

Corollary 3.5. *Under the assumptions* (A.1)-(A.2), *the inequalities*

$$W_n(\mathcal{D}) \leq_{ci} W_n, \qquad n = 0,1,...(3.27)$$

$$R_n(\mathcal{D}) \leq_{ci} R_n \qquad n = 0,1,...(3.28)$$

and

$$S_n(\mathcal{D}) \leq_{ci} S_n \qquad n = 0,1,...(3.29)$$

hold.

The next corollary parallels Corollary 3.2 to the present set-up and shows how some of the bounds could possibly be improved.

Corollary 3.6. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two σ-fields such that $\mathcal{D}_1 \subseteq \mathcal{D}_2 \subseteq \mathcal{F}$, which both satisfy* (A.2). *Under the finite mean assumption* (A.1), *the inequalities*

$$W_n(\mathcal{D}_1) \leq_{ci} W_n(\mathcal{D}_2) \leq_{ci} W_n, \qquad n = 0, 1, ... \quad (3.30)$$

and

$$S_n(\mathcal{D}_1) \leq_{ci} S_n(\mathcal{D}_2) \leq_{ci} S_n \qquad n = 0, 1, ... \quad (3.31)$$

hold.

Proof: The rightmost inequalities are already contained in Corollary 3.5. (with $\mathcal{D} = \mathcal{D}_2$). In order to prove the leftmost ones, fix $n = 0, 1, ...$ and consider any integrable $I\!R$-valued RV Y which is *conditionally independent* of the RV u_n given each one of the σ-fields $\mathcal{D}_1$ and $\mathcal{D}_2$, a condition which implies

$$E[u_n^k \cdot Y | \mathcal{D}_i] = E[u_n^k | \mathcal{D}_i] \cdot E[Y | \mathcal{D}_i], \quad i = 1, 2 \qquad (3.32)$$

for all $1 \leq k \leq K$. Any integrable $\mathcal{G}_n$-measurable RV is such a RV owing to the enforced condition (A.2) on both σ-fields $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{G}_n$ is the σ-field entering the definition of (A.2).

Under the assumed conditions on the RV Y, the inclusion $\mathcal{D}_1 \subseteq \mathcal{D}_2$ and (3.32) easily give

$$(u_n^k \cdot Y)^{\mathcal{D}_1} = \left((u_n^k \cdot Y)^{\mathcal{D}_2} \right)^{\mathcal{D}_1} = \left(u_n^{k,\mathcal{D}_2} \cdot Y^{\mathcal{D}_2} \right)^{\mathcal{D}_1} \qquad (3.33)$$

by an application of the smoothing property for conditional expectations. Comparison of (3.33) with (3.32) now implies

$$E[u_n^k | \mathcal{D}_1] \cdot E[Y | \mathcal{D}_1] = \left(u_n^{k,\mathcal{D}_2} \cdot Y^{\mathcal{D}_2} \right)^{\mathcal{D}_1} \qquad (3.34)$$

for all $1 \leq k \leq K$.

The RV σ_n being $\mathcal{G}_n$-measurable, it readily follows from (3.34) that for all $1 \leq k \leq K$,

$$u_n^{k,\mathcal{D}_1} \cdot \sigma_n^{k,\mathcal{D}_1} = \left(u_n^{k,\mathcal{D}_2} \cdot \sigma_n^{k,\mathcal{D}_2} \right)^{\mathcal{D}_1} \qquad (3.35)$$

and therefore

$$O_{n+1}^{k,\mathcal{D}_1} = u_n^{k,\mathcal{D}_1} \cdot \sigma_n^{k,\mathcal{D}_1} - \tau_{n+1}^{k,\mathcal{D}_1} = \left(O_{n+1}^{k,\mathcal{D}_2} \right)^{\mathcal{D}_1} \qquad (3.36)$$

since $(\tau_{n+1}^{k,\mathcal{D}_2})^{\mathcal{D}_1} = \tau_{n+1}^{k,\mathcal{D}_1}$ by the smoothing property.

The remarks following the proof of Theorem 3.3 now implies that the inequalities

$$W_n^k(\mathcal{D}_1) \leq E[W_n^k(\mathcal{D}_2)|\mathcal{D}_1] \tag{3.37}$$

and

$$S_n^k(\mathcal{D}_1) \leq E[S_n^k(\mathcal{D}_2)|\mathcal{D}_1] \tag{3.38}$$

hold, by translating (3.13) and (3.15) with $\mathcal{D} = \mathcal{D}_1$ and constituting sequence

$$\langle W, \sigma_n^{\mathcal{D}_2}, u_n^{\mathcal{D}_2}, \tau_{n+1}^{\mathcal{D}_2},\ n = 0, 1, ... \rangle. \tag{3.39}$$

The proof is completed by noting from Jensen's inequality that (3.37)-(3.38) are equivalent to the leftmost inequalities (3.30)-(3.31). □

It should be pointed out that the inequality

$$R_n(\mathcal{D}_1) \leq_{ci} R_n(\mathcal{D}_2) \qquad n = 0, 1, ... \tag{3.40}$$

may not hold in general under the assumptions of Corollary 3.6. Indeed, the remarks following the proof of Theorem 3.3 indicate that the a.s. equalities

$$E[u_n^{k,\mathcal{D}_2} \cdot W_n^k(\mathcal{D}_2)|\mathcal{D}_1] = u_n^{k,\mathcal{D}_1} \cdot E[W_n^k(\mathcal{D}_2)|\mathcal{D}_1], \qquad n = 0, 1, .. \tag{3.41}$$

for all $1 \leq k \leq K$, are sufficient to give (3.40). Unfortunately this does not seem to follow in general from (3.34). A sufficient condition to ensure (3.41) is that the projected constituting sequence (3.39) itself satisfies (A.2) with respect to the σ-field $\mathcal{D}_1$.

3.3 Steady-state analysis.

The bounds obtained thus far are transient in nature but extend readily to statistical (or steady-state) equilibrium when meaningful. To that end, consider the following additional conditions (A.3) and (A.3bis) on the constituting sequences (3.11) and (3.12) (for some σ-field $\mathcal{D}$), where

(A.3) The RV's $\{(\sigma_n, u_n, \tau_{n+1})\}_0^\infty$ form a *stationary ergodic* sequence; and

(A.3bis) The RV's $\{(\sigma_n^{\mathcal{D}}, u_n^{\mathcal{D}}, \tau_{n+1}^{\mathcal{D}})\}_0^\infty$ form a *stationary ergodic* sequence.

The next proposition is a vector version of the increasing scheme technique of Loynes [17]; its proof is given in Appendix A.

Theorem 3.7. *Under the assumptions* (A.1) *and* (A.3), *(resp.* (A.1) *and* (A.3bis)), *the condition*

$$E[u_n^k \cdot \sigma_n^k] < E[\tau_{n+1}^k], \quad 1 \leq k \leq K, \qquad n = 0, 1, ... \tag{3.42}$$

(resp.

$$E[u_n^{k,\mathcal{D}} \cdot \sigma_n^{k,\mathcal{D}}] < E[\tau_{n+1}^{k,\mathcal{D}}], \quad 1 \leq k \leq K), \qquad n = 0, 1, ... \quad (3.43)$$

ensures that the sequence of waiting times $\{W_n\}_0^\infty$ *(resp.* $\{W_n(\mathcal{D})\}_0^\infty$*) converges weakly to some non-defective probability distribution function on* IR_+^K.

The reader will readily check from (3.17) that under the assumptions (A.1)-(A.3bis), the queueing systems generated by the constituting sequences (3.11) and (3.12) exhibit the *same* stability condition, to wit (3.42). In the sequel, the IR_+^K-valued RV W_∞ (resp. $W_\infty(\mathcal{D})$) will be any RV on $(\Omega, \mathcal{F}, P)$ distributed according to the limiting distribution of $\{W_n\}_0^\infty$ (resp. $\{W_n(\mathcal{D})\}_0^\infty$), whose existence is guaranteed by Theorem 3.7 under the stability condition (3.42). Similar definitions hold for R_∞ (resp. $R_\infty(\mathcal{D})$), S_∞ (resp. $S_\infty(\mathcal{D})$) and T_∞^i (resp. $T_\infty^i(\mathcal{D})$), $i = 1, 2$, with obvious interpretations.

Theorem 3.8. *Under the assumptions* (A.1)-(A.3bis), *whenever the stability condition (3.42) is satisfied, the inequalities*

$$W_\infty(\mathcal{D}) \leq_{ci} W_\infty \quad (3.44)$$

$$R_\infty(\mathcal{D}) \leq_{ci} R_\infty \quad (3.45)$$

$$S_\infty(\mathcal{D}) \leq_{ci} S_\infty \quad (3.46)$$

and

$$T_\infty^i(\mathcal{D}) \leq_{ci} T_\infty^i, \quad i = 1, 2 \quad (3.47)$$

hold true.

Proof. A proof is available in Appendix B. □

4. THE FIRST FAMILY OF BOUNDS

Consider the K-dimensional FJ queue described in Section 2 under the synchronization condition (2.10), with assumption (A.1) enforced throughout. The discussion given in Section 3 clearly indicate how bounds can be obtained on the statistics of $\{W_n\}_0^\infty, \{R_n\}_0^\infty, \{S_n\}_0^\infty$ and $\{T_n^i\}_0^\infty$, $i = 1, 2$. This is done by appropriately selecting several sub σ-fields $\mathcal{D}$ of $\mathcal{F}$, with a view towards giving a different system interpretation in each case. The material of this section being illustrative of the methodology of Section 3, no attempts are made to give the strongest possible results.

4.1. The FJ queue with deterministic arrival and loading patterns

Consider the sub σ-algebra $\mathcal{D}_1$ of $\mathcal{F}$ given by

$$\mathcal{D}_1 = \sigma\{W, \sigma_n, \ n = 0, 1, ...\} \quad (4.1)$$

under the conditions (A.2) and (A.4), where

(A.4) The σ-field $\mathcal{D}_1$ is *independent* of the σ-field $\sigma\{u_n, \tau_{n+1},\ n = 0,1,...\}$.

This set of conditions will be satisfied for instance if the independence assumption (I) holds. Moreover, note also that *under* (A.4), condition (A.2) is equivalent to the RV u_n be independent of the RV's $\{u_k, \tau_{k+1},\ 0 \leq k < n\}$ for all $n = 0,1,...$

With the RV's $\{O_{n+1}\}_0^\infty$ defined by (3.18), condition (A.4) implies

$$O_{n+1}^{k,\mathcal{D}_1} = E[u_n^k] \cdot \sigma_n^k - E[\tau_{n+1}]. \qquad n = 0,1,...(4.2)$$

for all $1 \leq k \leq K$. A direct application of Theorem 3.3 then shows that the FJ queue with both *deterministic* input and loading patterns constitutes a *lower bound* system to the original one (in the convex increasing order sense). Interest in this FJ queueing system with deterministic components becomes apparent whenever the RV's W and $\{\sigma_n\}_0^\infty$ have independent components, so that the families of RV's $\{R_n^k(\mathcal{D}_1)\}_0^\infty$, $1 \leq k \leq K$, turn out to be *mutually independent*. A similar statement holds for the quantities $\{S_n^k(\mathcal{D}_1)\}_0^\infty$, $1 \leq k \leq K$.

This lower bound system receives a different interpretation depending on the situation. In case (C.1), it is clear that

$$E[u_n^k] = P[k \in I_n] \qquad n = 0,1,...(4.3)$$

for all $1 \leq k \leq K$, whereas in case (C.2), $E[u_n^k]$ represents the mean value of a common scaling factor, namely,

$$E[u_n^1] = \cdots = E[u_n^K] = E[v_n]. \qquad n = 0,1,...(4.4)$$

4.2 The FJ queue with deterministic service times

Consider the σ-field $\mathcal{D}_2$ defined by

$$\mathcal{D}_2 = \sigma\{W, \tau_{n+1},\ n = 0,1,...\} \qquad (4.5)$$

under the conditions (A.2) and (A.5), where

(A.5) the σ-field $\mathcal{D}_2$ is *independent* of the σ-field $\sigma\{\sigma_n, u_n,\ n = 0,1,...\}$.

Here again, this set of conditions will be satisfied under assumption (I), whereas under (A.5), the condition (A.2) is equivalent to the RV u_n be independent of the RV's σ_n and $\{u_k, \sigma_k,\ 0 \leq k < n\}$ for all $n = 0,1,...$ Under the assumptions (A.1)-(A.2) and (A.5), it is clear that

$$O_{n+1}^{k,\mathcal{D}_2} = E[u_n^k] \cdot E[\sigma_n^k] - \tau_{n+1}. \qquad n = 0,1,...(4.6)$$

and Theorerm 3.3 shows that the FJ queue with *deterministic* service times also constitutes a lower bound system to the original system. If the system is initially empty (i. e., $W = 0$) and the homogeneity condition

$$E[u_n^k] \cdot E[\sigma_n^k] = s^k \qquad n = 0, 1, ...(4.7)$$

holds for all $1 \leq k \leq K$, then the system response times $\{T_n^2(\mathcal{D}_2)\}_0^\infty$ can be expressed as the response times in a G/D/1 system with interarrival stream $\{\tau_{n+1}\}_0^\infty$ and constant service requirements $\max_{1 \leq k \leq K} s^k$. More precisely,

$$T_n^2(\mathcal{D}_2) = V_n + \max_{1 \leq k \leq K} s^k, \qquad n = 0, 1, ...(4.8)$$

where the RV's $\{V_n\}_0^\infty$ are the successive waiting times in a G/D/1 queue, given by the recursion

$$V_{n+1} = [V_n + \max_{1 \leq k \leq K} s^k - \tau_{n+1}]^+ \qquad n = 0, 1, ...(4.9)$$

with $V_0 = 0$.

4.3. The FJ queue with divisible statistics

Assume the synchronization condition (2.10) to hold, and the (synchronized) interarrival times $\{\tau_{n+1}\}_0^\infty$ to be K-divisible in the sense that the following conditions (D.1)-(D.3) are satisfied, where

(D.1) The IR_+-valued RV's $\{\tau_{n+1},\ n = 0, 1, ...\}$ are *mutually independent*;

(D.2) There exists a sequence $\{\tilde{\tau}_{n+1}\}_0^\infty$ of IR_+^K-valued RV's such that

$$\tau_{n+1} = \frac{1}{K} \sum_{k=1}^{K} \tilde{\tau}_{n+1}^k. \qquad n = 0, 1, ...(4.10)$$

where the K sequences $\{\tilde{\tau}_{n+1}^1\}_0^\infty, \cdots, \{\tilde{\tau}_{n+1}^K\}_0^\infty$ of IR_+-valued RV's are *mutually independent* and *identically distributed*; and

(D.3) The families of RV's $\{\tilde{\tau}_{n+1},\ n = 0, 1, ...\}$ and $\{W, \sigma_n, u_n,\ n = 0, 1, ...\}$ are *mutually independent*.

In the case of renewal processes, this will be satisfied if the distribution function of the RV's $\{\tau_{n+1}\}_0^\infty$ is K-divisible in the classical sense. Moreover, it is plain from (D.1)-(D.2) that for all $n = 0, 1, ...$, the RV's $\{\tilde{\tau}_{n+1}^1, \cdots, \tilde{\tau}_{n+1}^K\}$ are *exchangeable* given the σ-field $\mathcal{T}$, with the notation

$$\mathcal{T} := \sigma\{\tau_{n+1},\ n = 0, 1, ...\}, \qquad (4.11)$$

and therefore

$$E[\tilde{\tau}_{n+1}^1 | \mathcal{T}] = \cdots = E[\tilde{\tau}_{n+1}^K | \mathcal{T}] = \tau_{n+1} \qquad n = 0, 1, ...(4.12)$$

with the last equation following from (4.10).

Consider now the non-synchronized queueing system generated by the constituting sequence

$$\langle W, \sigma_n, u_n, \tilde{\tau}_{n+1},\ n = 0, 1, ... \rangle, \tag{4.13}$$

and observe under the assumptions (D.1)-(D.3) that the σ-field $\mathcal{D}_3$ defined by

$$\mathcal{D}_3 := \sigma\{W, \sigma_n, u_n, \tau_{n+1},\ n = 0, 1, ...\} \tag{4.14}$$

clearly satisfies the assumption (A.2) (with respect to (4.13)). In complete analogy with (3.18), if the RV's $\{\tilde{O}_{n+1}\}_0^\infty$ are defined componentwise by

$$\tilde{O}_{n+1}^k = u_n^k \cdot \sigma_n^k - \tilde{\tau}_{n+1}^k \qquad n = 0, 1, ... \tag{4.15}$$

for all $1 \leq k \leq K$, then the equality (4.12) combines to the independence assumption (D.3) in order to yield

$$\tilde{O}_{n+1}^{k,\mathcal{D}_3} = u_n^k \cdot \sigma_n^k - \tau_{n+1} \qquad n = 0, 1, ... \tag{4.16}$$

for all $1 \leq k \leq K$.

¿From the results of Section 3, the system generated by (4.13) thus provides an *upper bound* to the original system. Interest in the system generated by such a divisible input stream becomes clear when the components of the RV's W, $\{\sigma_n\}_0^\infty$ and $\{u_n\}_0^\infty$ are all mutually independent, since in that case the system (4.13) provides a *decoupled* upper bound system to the original FJ queue in the sense defined in Section 3. The interested reader is invited to consult [2,5] for a specific example under renewal type assumptions.

In the case (C.2), the decoupling mentioned earlier cannot take place since the components RV's $\{u_n^1, \cdots, u_n^K\}$ are correlated, being all equal to v_n for all $n = 0, 1, ...$ However, if the RV's $\{v_n\}_0^\infty$ are assumed K-divisible as defined above, it is possible to extend this type of upper bound by considering a system with constituting sequence

$$\langle W, \sigma_n, \tilde{u}_n, \tilde{\tau}_{n+1},\ n = 0, 1, ... \rangle \tag{4.17}$$

where the $I\!R_+^K$-valued RV's $\{\tilde{u}_n\}_0^\infty$ satisfy conditions analoguous to (D.1)-(D.3) with respect to the sequence $\{v_n\}_0^\infty$. This will not discussed any further in this paper for the sake of brevity. Another reason for not pursuing this line of thought comes from the fact that the upper bounds obtained in this section can be improved by the methods discussed in Section 5. There, statements will be

obtained in a stronger stochastic sense than in the convex increasing ordering used in this section, but only on the statistics of the system response times $\{T_n^i\}_0^\infty$, $i = 1,2$. As pointed out by the authors in [4], bounding systems such as (4.13) and (4.17) are the only ones for which the strong vector ordering statement of Corollary 3.5 holds.

4.4 Refinements

Let the σ-fields $\mathcal{D}_1'$ and $\mathcal{D}_2'$ be defined by

$$\mathcal{D}_i' := \mathcal{D}_i \vee \sigma\{u_n,\ n = 0,1,...\} \quad i = 1,2 \tag{4.18}$$

and note that they always trivially satisfy the condition (A.2). Consider the σ-field $\mathcal{D}_1'$ under the condition (A.4bis),where

(A.4bis) The σ-field $\mathcal{D}_1'$ is *independent* of the σ-field $\sigma\{\tau_{n+1},\ n = 0,1,...\}$.

In this set-up, the results of Section 4.1 also hold when replacing $\mathcal{D}_1$ by $\mathcal{D}_1'$. A similar conclusion is obtained when the σ-field $\mathcal{D}_2'$ is used, provided condition is enforced, where

(A.5bis) The σ-field $\mathcal{D}_2'$ is *independent* of the σ-field $\sigma\{\sigma_n,\ n = 0,1,...\}$.

Observe that the queueing system with constituting sequence projected on the σ-field $\mathcal{D}_1'$ is the one with constituting sequence

$$\langle W, \sigma_n, u_n, E[\tau_{n+1}],\ n = 0,1,...\rangle. \tag{4.19}$$

It is now easy to see that the σ-field $\mathcal{D}_1'$ satisfies condition (A.2) with respect to (4.19), for this is equivalent to the fact that the RV u_n is independent of the RV's $\{u_k, 0 \leq k < n\}$, a fact which was noted to hold under the conditions (A.2) and (A.4). Under the conditions (A.2) and (A.5), similar comments apply to the σ-field $\mathcal{D}_2$ with respect to the constituting sequence

$$\langle W, E[\sigma_n], u_n, \tau_{n+1},\ n = 0,1,...\rangle. \tag{4.20},$$

which is obtained from projecting (4.1) on the σ-field $\mathcal{D}_2'$.

¿From the obvious σ-field inclusions

$$\mathcal{D}_i \subseteq \mathcal{D}_i', \quad i = 1,2, \tag{4.21}$$

it follows by Corollary 3.6 (and the remarks following its proof) that for $i = 1,2$,

$$W_n(\mathcal{D}_i) \leq_{ci} W_n(\mathcal{D}_i') \leq_{ci} W_n, \qquad n = 0,1,... \tag{4.22}$$

$$R_n(\mathcal{D}_i) \leq_{ci} R_n(\mathcal{D}_i') \leq_{ci} R_n, \qquad n = 0, 1, ...(4.23)$$

and

$$S_n(\mathcal{D}_i) \leq_{ci} S_n(\mathcal{D}_i') \leq_{ci} S_n, \qquad n = 0, 1, ...(4.24)$$

provided the assumptions (A.1)-(A.2), (A.4)-(A.4bis), and (A.1)-(A.2), (A.5)-(A.5bis) hold, respectively.

Consider first the case $i = 1$. Equations (4.22)-(4.24) imply that the system where only the arrival pattern $\{\tau_n\}_1^\infty$ is made deterministic provides a better lower bound than the one where both arrival and loading patterns $\{\tau_n\}_1^\infty$ and $\{u_n\}_0^\infty$ are made deterministic. For instance, for model (C.1), the RV $T_n^1(\mathcal{D}_1')$ takes the form

$$T_n^1(\mathcal{D}_1') = \max_{k \in I_n} R_n^k(\mathcal{D}_1'). \qquad n = 0, 1, ...(4.25)$$

However, such a refined bound will only be computable if both the RV's $\{\sigma_n\}_0^\infty$ and $\{u_n\}_0^\infty$ have mutually independent components.

Similar results hold for $i = 2$. For the special case (C.2) under the synchronization condition (2.10), if the system is initially empty (i. e., $W = 0$) and the homogeneity condition

$$E[\sigma_n^k] = \sigma^k \qquad n = 0, 1, ...(4.26)$$

holds for all $1 \leq k \leq K$, then the system response times $\{T_n^2(\mathcal{D}_2')\}_0^\infty$ can be interpreted as the system response times in a G/G/1 system with interarrival stream $\{\tau_{n+1}\}_0^\infty$ and service requirements $\{v_n \cdot \max_{1 \leq k \leq K} \sigma^k\}_0^\infty$. More precisely,

$$T_n^2(\mathcal{D}_2') = V_n' + v_n \cdot \sigma^k \qquad n = 0, 1, ..., (4.26)$$

where the RV's V_n' are the successive waiting times in the G/G/1 queue defined by the recursion

$$V_{n+1}' = [V_n' + v_n \cdot \sigma^k - \tau_{n+1}]^+. \qquad n = 0, 1, ...(4.27)$$

with $V_0' = 0$.

4.5. A Counter Example

At this point in the discussion, the reader may entertain other possible extensions to the previous results. A natural idea consists in making deterministic only part of the K arrival streams, say $L (1 < L \leq K)$ of them while keeping the remaining $K - L$ streams unchanged. Under standard

independence and renewal assumptions, such an approach would lead to a partially decoupled system composed of a FJ queue of smaller dimension fed by the initial arrival process and of an independent FJ system with deterministic arrivals (and thus composed of independent channels). The exact solution being available for two dimensional FJ queues (at least in some particular cases [1,12]), the system response time statistics would also be computable for such a composite system when $L = K-2$. The reader might hope at first that such a partially decoupled system still provides a lower bound for the initial system since its constituting sequence is less variable. Although the simulation results of [5] might lend credence to such a conjecture, this is unfortunately not true as shown by the following counter example. Consider a two-queue FJ system with *deterministic* loading and service sequences satisfying

$$\sigma_n^1 = \sigma_n^2 = \sigma, \quad u_n^1 = u_n^2 = 1 \qquad n = 0, 1, ... \text{(4.28)}$$

and with nondeterministic synchronized interarrival times

$$\tau_n^1 = \tau_n^2 = \tau_n. \qquad n = 1, 2, ... \text{(4.29)}$$

Let $\{W_n\}_0^\infty$ be the sequence of waiting times in this system where $W_0 = W = 0$ and let $\{\tilde{W}_n\}_0^\infty$ be the corresponding sequence when the arrival sequences $\{\tilde{\tau}_{n+1}\}_0^\infty$ are given by $\tilde{\tau}_{n+1}^1 = E[\tau_{n+1}]$ and $\tilde{\tau}_{n+1}^2 = \tau_{n+1}$ for all $n = 0, 1, ...$

With an obvious notation, the system response time for the first incoming customer is given by

$$T_1 = \sigma + (\sigma - \tau_1)^+ \qquad (4.30)$$

and

$$\tilde{T}_1 = \sigma + (\sigma - \min(\tau_1, E[\tau_1]))^+, \qquad (4.32)$$

respectively. Hence, $\tilde{T}_1 \geq T_1$ and the event $[\tilde{T}_1 > T_1]$ has positive probability thus proving that $\tilde{T}_1 \geq_{st} T_1$ (where $\geq_{st}$ denotes the strong order on distribution functions [19,21]), and consequently the ordering $\tilde{T}_1 \geq_{cs} T_1$ holds, i. e., the new system is not a lower bound to the original system.

4.6 Steady-state analysis

Assume that the constituting sequence of the original FJ queue satisfies condition (A.3). For all instances of sub σ-fields considered in this section, condition (A.3bis) is an immediate consequence of condition (A.3). Hence, owing to Theorem 3.8, all the transient bounds derived so far can be given a steady-state version provided the initial constituting sequence fulfills the stability condition of Theorem 3.7.

4.7 Ross' conjecture

It was conjectured by Ross and proved in [20] and [6] that the response times in a G/G/1 queue in random environment are always larger for the convex ordering than the corresponding response times in G/G/1 queue where the environment is averaged out. This basic result can also be established for the generalized response times considered here. A proof of such a result can be obtained by combining the results of Section 4 to the methods of [6]; the proof is omitted for the sake of brevity.

5. A FAMILY OF UPPER BOUNDS

A second bounding methodology is developed in this section. Its application to the situation treated here leads very naturally to the definition of a family of queueing systems that provide upper bound on the performance measures of interest.

5.1 Associated RV's

This second bounding methodology is based in an essential way on the notion of associated RV's [7,10]. More concretely, the $I\!R$-valued RV's $\{X_1, ..., X_K\}$ are *associated* if and only if, with the notation $X := (X_1, ..., X_K)$, the inequality

$$E[f(X)g(X)] \geq E[f(X)]E[g(X)] \tag{5.1}$$

holds for all pair of *monotone non-decreasing* mappings $f, g : I\!R^K \to I\!R$ for which the expectations $E[f(X)], E[g(X)]$ and $E[f(X)g(X)]$ exist.

In order to explain the usefulness of this concept in the present context, it will be convenient to say that the $I\!R$-valued RV's $\{\overline{X}_1, ..., \overline{X}_K\}$ form *independent versions* of the RV's $\{X_1, ..., X_K\}$ if

(i) : The RV's $\{\overline{X}_1, ..., \overline{X}_K\}$ are *mutually independent*, and

(ii) : For every $1 \leq k \leq K$, the RV's X_k and $\overline{X}_k$ have the *same* probability distribution.

The following proposition is an easy consequence of the definition (5.1) [8,11].

Theorem 5.1. *If the RV's $\{X_1, ...X_K\}$ are associated, then the inequality*

$$P[\max_{1 \leq k \leq K} X_k \leq x] \geq P[\max_{1 \leq k \leq K} \overline{X}_k \leq x] \tag{5.2}$$

holds true for all x in $I\!R$.

This result can be viewed as a statement on the stochastic ordering between the maximum of the RV's $\{X_1, ..., X_K\}$ and the corresponding quantity for the independent versions. More precisely,

if X and Y are two IR-valued RV's defined on Ω, then the (distribution of the) RV X is said to be *greater* than the (distribution of the) RV Y in the *stochastic order* if and only if

$$P[Y > t] \leq P[X > t] \tag{5.3}$$

for all t in IR; this is denoted in short by $Y \leq_{st} X$. As well known [19, Prop. 8.1.2, p. 252], (5.3) is equivalent to the statement that

$$E[\phi(Y)] \leq E[\phi(X)] \tag{5.4}$$

for all *monotone non-decreasing* mappings $\phi : IR \to IR$ for which (5.4) makes sense. With this notation, Proposition 5.1 can be restated simply as saying that

$$\max_{1 \leq k \leq K} X_k \leq_{st} \max_{1 \leq k \leq K} \overline{X}_k. \tag{5.5}$$

The elements of a "calculus" for associated RV's are provided in [11, pp. 29-31]. Some of these facts, which are often used in the discussion, have been collected in the next lemma for easy reference.

Lemma 5.2.

(i): *Independent RV's are always associated;*

(ii): *The union of independent collections of associated RV's forms a set of associated RV's;*

(iii): *Any subset of a family of associated RV's forms a set of associated RV's, and*

(iv): *Any monotone non-decreasing function of associated RV's generates a set of associated RV's.*

The basic bounding result is given in the next proposition. Let $\{\overline{X}_1, ..., \overline{X}_K\}$ be *independent versions* of some IR-valued RV's $\{X_1, ..., X_K\}$. For any subset I of the index set $\{1, \cdots, K\}$, define the IR^K-valued RV X^I by posing

$$X_k^I = \begin{cases} \overline{X}_k & \text{if } k \in I; \\ X_k & \text{if } k \notin I. \end{cases} \tag{5.6}$$

Note that $X^I = X$ when $I = \emptyset$ and that $X^I = \overline{X}$ when $I = \{1, \cdots, K\}$. The next result constitutes a strengthening of Theorem 5.1

Theorem 5.3. *Assume the RV's $\{X_1, ..., X_K\}$ and their independent versions $\{\overline{X}_1, ..., \overline{X}_K\}$ to be mutually independent. If the RV's $\{X_1, ..., X_K\}$ are associated, then*

$$\max_{1 \leq k \leq K} X_k^I \leq_{st} \max_{1 \leq k \leq K} X_k^J. \tag{5.7}$$

for any pair I and J of subsets of $\{1,\cdots,K\}$ such that $I \subseteq J$.

Proof. Define the $I\!R$-valued RV $V^{I,J}$ to be

$$V^{I,J} := \max_{k \in J \setminus I} X_k^I = \max_{k \in J \setminus I} X_k^J, \tag{5.8}$$

and observe that

$$\max_{1 \le k \le K} X_k^I = \max\{V^{I,J}, \max_{k \in J \setminus I} X_k\} \tag{5.9}$$

while

$$\max_{1 \le k \le K} X_k^J = \max\{V^{I,J}, \max_{k \in J \setminus I} \overline{X}_k\}. \tag{5.10}$$

The RV's X and $\overline{X}$ being independent, the RV's $\{X_1,...,X_K,\overline{X}_1,...,\overline{X}_K\}$ are thus associated by Lemma 5.2 (i)-(ii), and so are also the RV's $\{V^{I,J},\max_{k\in J\setminus I} X_k\}$ by virtue of Lemma 5.2 (iv). Consequently, Theorem 5.1 immediately yields

$$\max_{k \in J \setminus I} X_k \le_{st} \max_{k \in J \setminus I} \overline{X}_k \tag{5.11a}$$

and

$$\max\{V^{I,J}, \max_{k \in J \setminus I} X_k\} \le_{st} \max\{\overline{V^{I,J}}, \overline{\max_{k \in J \setminus I} X_k}\} \tag{5.11b}$$

and upon combining (5.11a)-(5.11b), it follows that

$$\max\{V^{I,J}, \max_{k \in J \setminus I} X_k\} \le_{st} \max\{\overline{V^{I,J}}, \overline{\max_{k \in J \setminus I} \overline{X}_k}\} \tag{5.12}$$

The independence of the RV's X and $\overline{X}$ implies that the RV's $V^{I,J}$ and $\max_{k\in J\setminus I}\overline{X}_k$ are independent, whence

$$\max\{\overline{V^{I,J}}, \overline{\max_{k \in J \setminus I} \overline{X}_k}\} =_{st} \max\{V^{I,J}, \max_{k \in J \setminus I} \overline{X}_k\}. \tag{5.13}$$

Combining (5.12) and (5.13) readily gives (5.7). □

5.2 The upper bound systems

The results of the previous section, especially Theorem 5.3, suggest introducing the following family of queueing systems in order to generate upper bounds on the response time statistics for the queueing system generated by the constituting system

$$\langle W, \sigma_n, u_n, \tau_{n+1},\ n = 0,1,...\rangle. \tag{5.14}$$

Let $\overline{W}$, $\{\overline{\sigma}_n\}_0^\infty$, $\{\overline{u}_n\}_0^\infty$ and $\{\overline{\tau}_{n+1}\}_0^\infty$ be sequences of IR_+^K-valued RV's defined on the probability triple $(\Omega, \mathcal{F}, P)$ under the following assumptions (A.6)-(A.8), where

(A.6) The collections of RV's $\{W, (\sigma_n, u_n, \tau_{n+1}),\ n = 0,1,...\}$ and $\{\overline{W}, (\overline{\sigma}_n, \overline{u}_n, \overline{\tau}_{n+1}),\ n = 0,1,...\}$ are *independent*;

(A.7) The sequences of RV's $\{\overline{W}^k, (\overline{\sigma}_n^k, \overline{u}_n^k, \overline{\tau}_{n+1}^k),\ n = 0,1,...\}$, $1 \le k \le K$, are *mutually independent*; and

(A.8) For each $1 \le k \le K$, the sequence of RV's $\{\overline{W}^k, (\overline{\sigma}_n^k, \overline{u}_n^k, \overline{\tau}_{n+1}^k),\ n = 0,1,...\}$ is *statistically indistinguishable* from the original sequence $\{W^k, (\sigma_n^k, u_n^k, \tau_{n+1}^k),\ n = 0,1,...\}$.

In analogy with (5.6), if X is any of the IR_+^K-valued RV's $\{W, \sigma_n, u_n, \tau_{n+1},\ n = 0,1,...\}$, define the IR_+^K-valued RV X^I to be

$$X_k^I = \begin{cases} \overline{X}_k & \text{if } k \in I; \\ X_k & \text{if } k \notin I \end{cases} \tag{5.15}$$

for any subset I of the index set $\{1,...,K\}$.

With this notation, for any subset I of the index set $\{1,...,K\}$, consider the K servers queueing system generated by the constituting sequence

$$\langle W^I, \sigma_n^I, u_n^I, \tau_{n+1}^I,\ n = 0,1,...\rangle. \tag{5.16}$$

All the quantities of interest for this system are superscripted by I. In particular, the corresponding waiting times and response times form sequences of IR_+^K-valued RV's $\{W_n^I\}_0^\infty$, $\{R_n^I\}_0^\infty$ and $\{S_n^I\}_0^\infty$; the former is generated componentwise by the recursive scheme

$$W_{n+1}^{I,k} = [W_n^{I,k} + u_n^{I,k} \cdot \sigma_n^{I,k} - \tau_{n+1}^{I,k}]^+, 1 \le k \le K, \qquad n = 0,1,... \tag{5.17}$$

with $W_0^I = W^I$, whereas the latter are defined by

$$R_n^{I,k} = u_n^{I,k} \cdot (W_n^{I,k} + \sigma_n^{I,k}) \qquad n = 0,1,... \tag{5.18}$$

and

$$S_n^{I,k} = W_n^{I,k} + u_n^{I,k} \cdot \sigma_n^{I,k} \qquad n = 0,1,... \tag{5.19}$$

for all $1 \le k \le K$. The system response times $\{T_n^{i,I}\}_0^\infty$, $i = 1,2$, are then given simply by

$$T_n^{1,I} = \max_{1 \le k \le K} R_n^{I,k} \qquad n = 0,1,... \tag{5.20}$$

and

$$T_n^{2,I} = \max_{1 \le k \le K} S_n^{I,k}. \qquad n = 0, 1, ...(5.21)$$

This system clearly reduces to the original system when $I = \emptyset$.

The definitions (5.17)-(5.19) of the RV's W_n^I, R_n^I and S_n^I are consistent with the definition (5.15) where the RV's $\overline{W}_n$, $\overline{R}_n$ and $\overline{S}_n$ are defined through (5.17)-(5.19) with $I = \{1, \cdots, K\}$. More specifically, the RV's $\{\overline{W}_n\}_0^\infty$ are generated componentwise by the recursive scheme

$$\overline{W}_{n+1}^k = [\overline{W}_n^k + \overline{u}_n^k \cdot \overline{\sigma}_n^k - \overline{\tau}_{n+1}^k]^+, 1 \le k \le K, \qquad n = 0, 1, ...(5.22)$$

with $W_0 = \overline{W}$, whereas the RV's $\{\overline{R}_n\}_0^\infty$ and $\{\overline{S}_n\}_0^\infty$ are defined by

$$\overline{R}_n^k = \overline{u}_n^k \cdot (\overline{W}_n^k + \overline{\sigma}_n^k) \qquad n = 0, 1, ...(5.23)$$

and

$$\overline{S}_n^k = \overline{W}_n^k + \overline{u}_n^k \cdot \overline{\sigma}_n^k \qquad n = 0, 1, ...(5.24)$$

for all $1 \le k \le K$.

5.3 The main stochastic ordering theorem

The material of the Sections 5.1 and 5.2 is now combined to obtain the following basic result for the systems of interest here.

Theorem 5.4. *Assume the RV's $\{R_n^1, \cdots, R_n^K\}$ and $\{S_n^1, \cdots, S_n^K\}$ to both form sets of associated RV's for all $n = 0, 1, ...$ Under the assumptions* (A.6)-(A.8), *for any two subsets I and J of the index set $\{1, \cdots, K\}$ such that $I \subseteq J$, the inequalities*

$$T_n^1 \le_{st} T_n^{1,I} \le_{st} T_n^{1,J} \qquad n = 0, 1...(5.25)$$

and

$$T_n^2 \le_{st} T_n^{2,I} \le_{st} T_n^{2,J} \qquad n = 0, 1...(5.26)$$

hold.

Proof. Only the righmost inequalities in (5.25)-(5.26) need to be established since the leftmost ones follow immediately from them upon taking I and J to be $\emptyset$ and I, respectively. The reader will observe that for each $n = 0, 1, ...$, the RV's $\{\overline{R}_n^1, \cdots, \overline{R}_n^K\}$ and $\{\overline{S}_n^1, \cdots, \overline{S}_n^K\}$ constitute *independent versions* for the RV's $\{R_n^1, \cdots, R_n^K\}$ and $\{S_n^1, \cdots, S_n^K\}$, respectively, as a result of the

assumptions (A.6)-(A.8). Moreover, by virtue of (A.6), the RV's $\{R_n^1, \cdots, R_n^K\}$ and $\{\overline{R}_n^1, \cdots, \overline{R}_n^K\}$ are independent, and so are $\{S_n^1, \cdots, S_n^K\}$ and $\{\overline{S}_n^1, \cdots, \overline{S}_n^K\}$. The hypotheses of Theorem 5.3 are thus satisfied and upon immediate identification, the result (5.25)-(5.26) follows. □

Sufficient conditions are now given to ensure that the collections of RV's $\{R_n^1, \cdots, R_n^K\}$ and $\{S_n^1, \cdots, S_n^K\}$ form sets of associated RV's for all $n = 0, 1, \ldots$. To that end, consider the assumptions (A.9)-(A.10), where

(A.9) The RV's $\{W, (\sigma_n, u_n, \tau_{n+1}),\ n = 0, 1, \ldots\}$ are *mutually independent*; and

(A.10) Each one of the collections of RV's $\{W^1, \cdots, W^K\}$ and $\{\sigma_n^1, u_n^1, -\tau_{n+1}^1, \cdots, \sigma_n^K, u_n^K, -\tau_{n+1}^K\}$, $n = 0, 1, \ldots$, form a set of *associated* RV's.

Theorem 5.5. *Under the assumptions* (A.9) *and* (A.10), *for all* $n = 0, 1, \ldots$ *each one of the three collections of RV's* $\{W_n^1, \cdots, W_n^K\}$, $\{R_n^1, \cdots, R_n^K\}$ *and* $\{S_n^1, \cdots, S_n^K\}$ *forms a set of associated RV's.*

This result was already obtained by Nelson and Tantawi [17, Appendix] under stronger assumptions and only for the case $I = \emptyset$. Their proof, used here, is an inductive one and applies with minor modifications to the more general situation.

Proof. The proof proceeds by induction. Take as induction hypothesis that the RV's $\{W_m^1, \cdots, W_m^K\}$ are *associated* for *some* $n = m \geq 0$. By virtue of (A.9), the RV W_m is independent of the RV $(\sigma_m, u_m, -\tau_{m+1})$, and therefore the induction hypothesis and (A.10) imply that the RV's $\{W_m^1, \cdots, W_m^K, \sigma_m^1, u_m^1, -\tau_{m+1}^1, \cdots, \sigma_m^K, u_m^K, -\tau_{m+1}^K\}$ are associated, upon applying Lemma 5.2 (ii). Part (iv) of Lemma 5.2 now gives the conclusion that the sets of RV's $\{W_n^1, \cdots, W_n^K\}$, $\{R_n^1, \cdots, R_n^K\}$ and $\{S_n^1, \cdots, S_n^K\}$ are three sets of associated RV's. In passing, this shows that the induction hypothesis holds for $n = m + 1$, and since it holds for $n = 0$, by virtue of assumption (A.10), it holds for all $n = 0, 1, \ldots$. □

The main result of this section is now obtained upon combining Theorems 5.4 and 5.5.

Theorem 5.6. *Under the assumptions* (A.6)-(A.10), *for any pair of subsets* I *and* J *of the index set* $\{1, \ldots, K\}$ *such that* $I \subseteq J$, *the inequalities*

$$T_n^i \leq_{st} T_n^{i,I} \leq_{st} T_n^{i,J}, \quad i = 1, 2 \qquad n = 0, 1, \ldots (5.27)$$

hold true.

The following corollary follows by a simple application of (5.4) and (5.27) (with $\phi(x) = x$).

Corollary 5.7. *Under the assumptions* (A.6)-(A.10), *the inequalities*

$$E[T_n^i] \leq E[T_n^{i,I}] \leq E[T_n^{i,J}], \quad i = 1, 2 \qquad n = 0, 1, \ldots (5.28)$$

hold true.

In the case $u_n^1 = \cdots = u_n^K = 1$ for all $n = 0, 1, ...$, Nelson and Tantawi [17] gave (5.28) with $I = J = \{1, ..., K\}$. Theorem 5.6 and its Corollary 5.7 thus provide a strengthened version of these earlier results.

5.4 A special case

Consider the FJ queue system of Section 2 with synchronized arrivals (2.10) and synchronized loading factors (2.3) of model (C.2), and assume the constituting sequence (5.14) to satisfy the *strong independence* assumption (Ibis). Conditions (A.9)-(A.10) are then both satisfied, the latter as a result of Lemma 5.2 (i), and the FJ queue system with constituting sequence (5.16) and $I = \{1, \cdots, K\}$ is constituted by K *independent* components, each one having the same statistics as the corresponding component in the original system. Theorem 5.5 thus implies that this fully decoupled system provides a *computable* upper bound to this initial system, in the sense that

$$T_n^2 \leq_{st} T_n^{2,\{1,...,K\}}, \qquad n = 0, 1, ... \text{(5.29)}$$

an equation obtained by specializing the leftmost inequality of (5.27) to the case $I = \{1, ..., K\}$.

Theorems 5.4 and 5.5 can also be used to generate better upper bounds on the performance measures of interest, which are still computable. To see this, assume the response time distribution to be computable for the synchronized model under consideration in L dimensions for some $2 \leq L \leq K$. Then for any subset I of $\{1, ..., K\}$ with cardinality L, (5.27) gives the bound

$$T_n^2 \leq_{st} T_n^{2,I} \leq_{st} T_n^{2,\{1,...,K\}}. \qquad n = 0, 1, ... \text{(5.30)}$$

which improves on (5.29). The statistics of the response times $\{T_n^{2,I}\}_0^\infty$ are computable since the corresponding system has $L+1$ *independent* components, namely, one L-dimensional synchronized FJ system and $K - L$ independent unsynchronized single server queues.

Theorem 5.4 can also be used directly to generate yet better upper bounds as follows. Let $\{I_1, \cdots, I_p\}$ be a partition of $\{1, \cdots, K\}$, i. e., $I_j \cap I_\ell = \emptyset$ for $1 \leq j < \ell \leq p$ and $\bigcup_{j=1}^p I_j = \{1, ..., K\}$, such that $|I_j| \leq L$ for $1 \leq j \leq p$. Define the $I\!R$-valued RV's $\{X_n^j\}_0^\infty$, $1 \leq j \leq p$, by

$$X_n^j := \max_{k \in I_j} S_n^k \qquad n = 0, 1, ... \text{(5.31)}$$

for all $1 \leq j \leq p$.

For each $n = 0, 1, ...$, the RV's $\{S_n^1, \cdots, S_n^K\}$ are associated by Theorem 5.5 and so are the RV's $\{X_n^1, \cdots, X_n^p\}$ by virtue of Lemma 5.2 (iv), whence

$$T_n^2 = \max_{1 \le j \le p} X_n^j \le_{st} \max_{1 \le j \le p} \overline{X}_n^j, \qquad n = 0, 1, ... (5.32)$$

where the RV's $\{\overline{X}_n^1, ..., \overline{X}_n^p\}$ form independent versions of $\{X_n^1, ..., X_n^p\}$. With this notation, it is also clear that

$$\overline{X}_n^j \le_{st} \max_{k \in I_j} \overline{S}_n^k \qquad n = 0, 1, ... (5.33)$$

for all $1 \le j \le p$ by a direct application of Theorem 5.6 to the FJ queue system made up of servers whose indices are in I_j. If the RV's $\{\Theta_n^{I_1,...,I_p}\}_0^\infty$ are now given by

$$\Theta_n^{I_1,...,I_p} := \max_{1 \le j \le p} \overline{X}_n^j, \qquad n = 0, 1, ... (5.34)$$

then the inequalities

$$T_n^2 \le_{st} \Theta_n^{I_1,...,I_p} \le_{st} T_n^{2,\{1,...,K\}}, \qquad n = 0, 1, ... (5.35)$$

are obtained by combining (5.32) and (5.33). The system with response times $\{\Theta_n^{I_1,...,I_p}\}_0^\infty$ thus provides a refinement on the upper bound (5.29). It is also computable since it is composed of p *independent* FJ systems, all of dimension no greater than L.

A similar discussion can be carried out in the context of model (C.1) under the independence assumption (I) provided the independent RV's σ_n, u_n and $-\tau_{n+1}$ have *associated* components for all $n = 0, 1, ...$

5.5 Steady State Analysis

Assume the constituting RV's (5.14) that define the original system to satisfy both the conditions (A.3) and (A.9). Consequently the RV's $\{\sigma_n\}_0^\infty$, $\{u_n\}_0^\infty$ and $\{\tau_{n+1}\}_0^\infty$ form three mutually independent sequences of i.i.d. $I\!R_+^K$-valued RV's. It is now easy to see that conditions (A.6)-(A.8) implies that the constituting sequences (5.16) also satisfies (A.3) for every subset I of the index set $\{1, \cdots, K\}$, i. e., both (5.14) and (5.16) satisfy the stability condition (3.42) at the same time. Therefore, if the stability condition (3.42) is enforced, Theorem 3.7 ensures that the sequence of RV's $\{W_n^I\}_0^\infty$ (for each subset I of $\{1, \cdots, K\}$) converges weakly to some non-defective distribution function on $I\!R_+^K$. Again, generic RV's which are distributed according to the limiting distribution functions of $\{W_n^I\}_0^\infty, \{R_n^I\}_0^\infty, \{S_n^I\}_0^\infty$ and $\{T_n^{i,I}\}_0^\infty$, $i = 1, 2$, are denoted simply by $W_\infty^I, R_\infty^I, S_\infty^I$

and $T_\infty^{i,I}$, respectively. The following proposition shows that the transient bounds of Theorem 5.6 also hold in statistical equilibrium.

Theorem 5.6. *Under the assumptions* (A.3) *and* (A.6)-(A.10), *whenever the stability condition (3.42) holds, the inequalities*

$$T_\infty^i \leq_{st} T_\infty^{i,I} \leq_{st} T_\infty^{i,J}, \quad i = 1,2 \tag{5.36}$$

hold for every pair of subsets I and J of the index set $\{1,\cdots,K\}$ such that $I \subseteq J$.

Proof. The proof follows from Theorem 5.5 and from the stability of the stochastic ordering $\leq_{st}$ under weak limits [21, Prop. 1.2.3, p. 6]. □

6. THE RENEWAL CASE - COMPUTABLE BOUNDS

This section is devoted to explicit calculations for some of the bounds obtained thus far. More specifically, the discussion is carried out when the arrivals are synchronized in the sense of (2.10), under the following set of *renewal* assumptions (R.1)-(R.6), where

(R.1) The RV W and the sequences of RV's $\{\sigma_n\}_0^\infty$, $\{u_n\}_0^\infty$ and $\{\tau_{n+1}\}_0^\infty$ are *mutually independent*;

(R.2) The RV's $\{\tau_{n+1}\}_0^\infty$ form a *renewal* sequence with common probability distribution function $A(\cdot)$;

(R.3) The sequences $\{\sigma_n^k\}_0^\infty$, $1 \leq k \leq K$, are *mutually independent*;

(R.4) For each $1 \leq k \leq K$, the RV's $\{\sigma_n^k\}_0^\infty$ form a *renewal* sequence with common probability distribution function $B_k(\cdot)$;

(R.5) The RV's $\{u_n,\ n = 0,1,...\}$ are *mutually independent*; and

(R.6) For each $1 \leq k \leq K$, the RV's $\{u_n^k\}_0^\infty$ form a sequence of *i.i.d.* RV's with common probability distribution function $H_k(\cdot)$.

Under these assumptions, the RV's $\{u_n^1,\cdots,u_n^K\}$ are not necessarily independent; moreover, it should also be clear that the assumptions (R.1)-(R.6) imply the independence assumption (I). To fix the notation, pose

$$E[u_n^k] := \nu_k \qquad n = 0,1,...(6.1)$$

for all $1 \leq k \leq K$, and as usual, define the arrival and service rates through the relations

$$\frac{1}{\lambda} := E[\tau_{n+1}] = \int_0^\infty t dA(t) \qquad n = 0,1,...(6.2)$$

and

$$\frac{1}{\mu_k} := E[\sigma_n^k] = \int_0^\infty t dB_k(t) \qquad n = 0, 1, ...(6.3)$$

for all $1 \leq k \leq K$, respectively.

It will be convenient to refer to the ***homogeneous*** situation as the one where the probability distribution functions $\{B_k(\cdot)\}_1^K$ and $\{H_k(\cdot)\}_1^K$ all coincide with some probability distribution functions $B(\cdot)$ and $H(\cdot)$, respectively, in which case, pose

$$\mu_1 = \cdots = \mu_K =: \mu \quad \text{and} \quad \nu_1 = \cdots = \nu_K =: \nu. \tag{6.4}$$

6.1 Computable bounds

As indicated in earlier sections, ***computable*** bounds are obtained whenever ***statistical decoupling*** takes place between the various components in the corresponding bounding systems. This follows from the elementary facts that for any set of ***independent*** $I\!R_+$-valued RV's $\{X_1, \cdots, X_K\}$,

$$P\left[\max_{1 \leq k \leq K} X_k \leq x\right] = \prod_{k=1}^K P[X_k \leq x] \tag{6.5}$$

for all $x \geq 0$ and

$$E\left[\max_{1 \leq k \leq K} X_k\right] = \int_0^\infty (1 - \prod_{k=1}^K P[X_k \leq x]) dx. \tag{6.6}$$

The first example is obtained from the results on convex ordering given in Section 4.1. Under the assumptions (R.1)-(R.5), the σ-field $\mathcal{D}_1$ defined by (4.1) satisfies both assumptions (A.2) and (A.4). Here, the definition (4.2) of the RV's $\{O_{n+1}^{k,\mathcal{D}_1}\}_0^\infty$ reduces to

$$O_{n+1}^{k,\mathcal{D}_1} = \nu_k \cdot \sigma_k^n - \frac{1}{\lambda} \qquad n = 0, 1, ...(6.7)$$

for all $1 \leq k \leq K$. The RV's $\{S_n^1(\mathcal{D}_1), \cdots, S_n^K(\mathcal{D}_1)\}$ are ***mutually independent***, and the remark (6.6) thus yields the expression

$$E[T_n^2(\mathcal{D}_1)] = \int_0^\infty (1 - \prod_{k=1}^K P[S_n^k(\mathcal{D}_1) \leq x]) dx. \qquad n = 0, 1, ...(6.8)$$

Note that for every $1 \leq k \leq K$, the RV's $\{S_n^k(\mathcal{D}_1)\}_0^\infty$ are the successive response times in a standard GI/GI/1 queue with deterministic arrival times $\{\frac{n}{\lambda}\}_0^\infty$ and service requirements $\{\nu_k \cdot \sigma_n^k\}_0^\infty$. Similar

comments can be made concerning the computation of the statistics of the system response times $\{T_n^1(\mathcal{D}_1)\}_0^\infty$. Although a similar discussion could be carried out for the refinement to the bounds made in Section 4.4, this will not be pursued here.

Consider now the upper bounds derived in Section 5. Take $I = \{1, ..., K\}$ and assume the RV's $\{u_n^1, \cdots, u_n^K\}$ to be *associated*. It is easy to check that the assumptions (A.9)-(A.10) are satisfied. By construction, the bounding system with constituting sequence (5.16) described by (5.22)-(5.24) exhibits *independent* components, and therefore

$$E[T_n^{1,\{1,...,K\}}] = \int_{0-}^{\infty} (1 - \prod_{k=1}^{K} P[u_n^k \cdot (W_n^k + \sigma_n^k) \leq x])dx. \qquad n = 0, 1, ...(6.9)$$

Here the RV's $\{W_n^k\}_0^\infty$ are defined for the original FJ queueing system through (2.5), and correspond to the successive waiting times in a GI/GI/1 system with arrival stream $\{\tau_{n+1}\}_0^\infty$ and service requirements $\{u_n^k \cdot \sigma_n^k\}_0^\infty$. This last expression (6.9) simplifies somewhat when each loading sequence $\{u_n^k\}_0^\infty$ is a $\{0,1\}$-valued *Bernoulli* sequence, since then

$$P[u_n^k \cdot (W_n^k + \sigma_n^k) \leq x] = \pi_k \cdot P[W_n^k + \sigma_n^k \leq x] + (1 - \pi_k) \qquad n = 0, 1, ...(6.10)$$

for all $x \geq 0$, with the notation

$$\pi_k = P[u_n^k = 1] \qquad n = 0, 1, ...(6.11)$$

for all $1 \leq k \leq K$.

Note that the bounding systems discussed above all exhibit the same stability condition, to wit

$$\max_{1 \leq k \leq K} \frac{\lambda \cdot \nu_k}{\mu_k} < 1. \qquad (6.12)$$

Under this condition, the formulas (6.8) and (6.9) readily extend to statistical equilibrium, with n replaced everywhere by ∞.

Under the renewal and independence assumptions stated above, more explicit expressions can be obtained for the steady-state versions of these bounds when the service time distributions $B_k(\cdot), 1 \leq k \leq K$, are all *exponential*. In the next two sections, the calculations are carried out for this special case, with (6.8) and (6.9) as point of departure for the lower and upper bounds, respectively. It will be convenient to use the notation $\{H_K\}_1^\infty$ for the partial sums of the harmonic series, i. e.,

$$H_K = \sum_{k=1}^{K} \frac{1}{k}. \qquad K = 1, 2, ...(6.13)$$

6.2 Lower bounds - exponential servers

As pointed out earlier in the remark following (6.8), the evaluation of $E[T^2_\infty(\mathcal{D}_1)]$ amounts to computing the equilibrium response time distribution for K independent single server systems; the k-th such system is a D/M/1 queue with arrival times $\{\frac{n}{\lambda}\}_0^\infty$ and exponential service times with parameter $\gamma_k = \mu_k/\nu_k, 1 \le k \le K$. It is well known [10] that the response time in such a D/M/1 system is exponentially distributed. More precisely, for all $1 \le k \le K$,

$$P[S^k_\infty(\mathcal{D}_1) > x] = e^{-\delta_k x}, \quad x \ge 0 \tag{6.14}$$

with

$$\delta_k := \gamma_k \cdot (1 - \beta_k) = \frac{\mu_k}{\nu_k} \cdot (1 - \beta_k), \tag{6.15}$$

where β_k is the smallest positive solution to the equation

$$\beta = \exp(-\frac{\gamma_k(1-\beta)}{\lambda}), \quad \beta \ge 0. \tag{6.16}$$

The expression (6.8) in statistical equilibrium now becomes

$$E[T^2_\infty(\mathcal{D}_1)] = \int_0^\infty (1 - \prod_{k=1}^K (1 - e^{-\delta_k x}))dx. \tag{6.17}$$

Elementary caculations now show that for all $x \ge 0$,

$$1 - \prod_{k=1}^K (1 - e^{-\delta_k x}) = \sum_{k=1}^K (-1)^{k+1} \sum_{I \in I_k} \exp(-\sum_{k \in I} \delta_k x) \tag{6.18}$$

with the simplifying notation

$$I_k := \{I \subseteq \{1, \cdots, K\} : |I| = k\}, \quad 1 \le k \le K. \tag{6.19}$$

For any non-empty subset I of $\{1, \cdots, K\}$, it is plain that

$$\int_0^\infty \exp(-\sum_{k \in I} \delta_k x)dx = (\sum_{k \in I} \delta_k)^{-1} \tag{6.20}$$

and direct substitution of (6.18)-(6.20) into (6.17) readily yields

$$E[T^2_\infty(\mathcal{D}_1)] = \sum_{k=1}^K (-1)^{k+1} \sum_{I \in I_k} (\sum_{k \in I} \delta_k)^{-1}. \tag{6.21}$$

In the *homogeneous* case, $\delta_k = \delta$ for all $1 \leq k \leq K$ and the easy identity

$$\sum_{k=1}^{K} \frac{(-1)^{k+1}}{k} \binom{K}{k} = \sum_{k=1}^{K} \frac{1}{k} \tag{6.22}$$

allows a rewriting of (6.23) in the following simpler form

$$E[T_\infty^2(\mathcal{D}_1)] = \frac{1}{\delta} \cdot \sum_{k=1}^{K} \frac{1}{k} = \frac{1}{\delta} H_K \tag{6.23}$$

since $|\mathcal{I}_k| = \binom{K}{k}$ for all $1 \leq k \leq K$.

6.3 Upper bounds - exponential servers

Under the foregoing assumptions, the computation of (6.9) in equilibrium passes through the calculation of the equilibrium response time distribution for K independent GI/GI/1 systems, which are GI/M/1 systems with Bernoulli loading as considered in Appendix C. For each one of these GI/M/1 queues with Bernoulli loading, the interarrival time distribution is still $A(\cdot)$ as specified by (R.2).

Following the discussion of Appendix C, for all $1 \leq k \leq K$, let α_k denote the smallest positive solution of the equation

$$\alpha = \tilde{A}_k^*(\mu_k(1-\alpha)), \quad \alpha \geq 0 \tag{6.24}$$

where

$$\tilde{A}_k^*(s) = \frac{\pi_k \cdot A^*(s)}{1 - (1-\pi_k) \cdot A^*(s)}, \quad s \geq 0, \tag{6.25}$$

and define the constant θ_k by

$$\theta_k = \mu_k \cdot (1 - \alpha_k). \tag{6.26}$$

By elementary calculations, it follows from (C.26) that

$$\begin{aligned} E[e^{-sW_\infty^k}]B^*(s) &= \left[(1-\alpha_k) + \alpha_k \cdot \frac{\theta_k}{\theta_k + s}\right] \cdot \left[\frac{\mu_k}{\mu_k + s}\right] \\ &= \left[\frac{\theta_k}{\theta_k + s}\right] \end{aligned} \tag{6.27}$$

for all $s \geq 0$, as expected. It is now plain from (6.10) and (6.27) that in statistical equilibrium, the expression (6.9) becomes

$$E[T_\infty^{1,\{1,\ldots,K\}}] = \int_0^\infty (1 - \prod_{k=1}^{K} \left[\pi_k \cdot (1 - e^{-\theta_k x}) + (1-\pi_k)\right]) dx. \tag{6.28}$$

In the *homogeneous case*, $\theta_k = \theta$ and $\pi_k = \pi$ for all $1 \leq k \leq K$, and (6.28) reduces to

$$E[T_\infty^{1,\{1,\ldots,K\}}] = \int_0^\infty (1 - (1 - \pi e^{-\theta x})^K) dx. \tag{6.29}$$

It is clear that

$$\begin{aligned} E[T_\infty^{1,\{1,\ldots,K\}}] &= \int_0^\infty \sum_{k=1}^{K} (-1)^{k+1} \binom{K}{k} \pi^k e^{-k\theta x}\, dx \\ &= \frac{1}{\theta} \cdot \sum_{k=1}^{K} (-1)^{k+1} \binom{K}{k} \frac{\pi^k}{k} \end{aligned} \tag{6.30}$$

by elementary calculations and (6.20). From this last expression and from the fact that $\int_0^\pi x^{k-1} dx = \frac{\pi^k}{k}$ for all $1 \leq k \leq K$, the reader can readily check that

$$\begin{aligned} E[T_\infty^{1,\{1,\ldots,K\}}] &= -\frac{1}{\theta} \int_0^\pi \frac{1}{x} \sum_{k=1}^{K} \binom{K}{k} (-x)^k dx \\ &= \frac{1}{\theta} \int_0^\pi \frac{(1-x)^K - 1}{(1-x) - 1} dx \\ &= \frac{1}{\theta} \int_0^\pi \sum_{k=0}^{K-1} (1-x)^k dx = \frac{1}{\theta} \sum_{k=1}^{K} \frac{1 - (1-\pi)^k}{k} \end{aligned} \tag{6.31}$$

where elementary properties of geometric series were used to obtain the prior to last equality.

7. ASYMPTOTIC ANALYSIS FOR HOMOGENEOUS FJ QUEUES

The derivation of asymptotics is now considered for a class of *homogeneous* FJ queues, as K, the number of servers, grows large. For sake of simplicity, the discussion is carried in statistical equilibrium, under the renewal assumptions (R.1)-(R.6) and the addititonal assumptions (R.7)-(R.8), where

(R.7) The RV's $\{\sigma_n^k,\ 1 \leq k \leq K,\ n = 0, 1, \ldots\}$ form a collection of *i.i.d.* RV's whose common distribution $B(\cdot)$ has a *rational* Laplace-Stieltjes transform.

(R.8) The loading RV's $\{u_n\}_0^\infty$ are given by

$$u_n^1 = \cdots = u_n^K = 1. \qquad n = 0, 1, \ldots \tag{7.1}$$

7.1 Asymptotics for GI/GI/1 systems

Consider a *stable* GI/GI/1 queueing system where $A(\cdot)$ and $B(\cdot)$ denote the probability distribution functions of the interarrival and service times, respectively. The Laplace-Stieltjes transform $B^*(\cdot)$ is assumed *rational* so that the function $s \to f(s)$ which is initially defined for $\Re(s) = 0$ by

$$f(s) = A^*(s)B^*(-s), \tag{7.2}$$

is *continuable* in the region $\Re(s) \geq 0$. Define μ^+ as

$$\mu^+ := \inf \{s \in I\!R_+ : f(s) < \infty\} \tag{7.3}$$

It is plain under the enforced assumptions that $\mu^+ > 0$ and $f(\mu^+) = \infty$. Consequently, the queueing system being stable, $f'(0) < 0$ and convexity of $f(\cdot)$ implies the existence of a unique real number q in $(0, \mu^+)$ such that

$$f(q) = 1. \tag{7.4}$$

Let W, R and I be generic RV's which are distributed respectively according to the stationary waiting time, response time and idle period distributions of the GI/GI/1 queue under consideration. The following result is available in the monograph by Borovkov [9, Thm. 11, p. 129] and is given here in appropriate form for handy reference.

Lemma 7.1 *The Laplace-Stieltjes transform of W is given by the expression*

$$E[e^{-sW}] = \frac{1 - P[W > 0]}{1 - \phi(s)} \tag{7.5}$$

where $\phi(\cdot)$ is a completely monotone function which is analytic in the region $\{s : \Re(s) > -(q + \epsilon)\}$ for some $\epsilon > 0$ and which satisfies the conditions

$$\phi(-q) = 1 \quad \text{and} \quad \phi'(-q) = \frac{f'(q)}{E[e^{-qI}] - 1}. \tag{7.6}$$

The Laplace-Stieltjes transform of the response time R is thus given by

$$E[e^{-sR}] = B^*(s)E[e^{-sW}] \tag{7.7}$$

and the function $x \to P[R > x]$, the so-called *complementary* distribution function of R, has Laplace transform given by

$$\int_0^\infty P[R > x]e^{-sx}dx = \frac{1}{s}\Big[1 - B^*(s)\frac{1 - P[W > 0]}{1 - \phi(s)}\Big]. \tag{7.8}$$

This transform function is analytic in the region $IR(s) > -(q+\epsilon)$ for $\epsilon > 0$ sufficiently small but for a pole of order 1 at $s = -q$, owing to the fact that $q < \mu^+$. The residue C associated to this pole is therefore given by

$$C = \frac{B^*(-q)(1 - E(e^{-qI}))}{qf'(q)}(1 - P[W > 0]) \tag{7.9}$$

and classical results on the first left singularity of Laplace transforms then yield the following estimate.

Lemma 7.2. *Under the foregoing assumptions,*

$$P[R > x] = Ce^{-qx}(1 + o(1)) \tag{7.10}$$

when x goes to infinity, with q and C given by (7.4) and (7.9), respectively.

7.2 Maximum of identically distributed RV's

The derivation of asymptotic bounds relies in an essential way on results of Lai and Robbins [15] on the asymptotic behavior of the maximum of *identically* distributed RV's. Let $\{Y_k\}_1^\infty$ be a family of *identically* distributed IR^+-valued RV's with common probability distribution function $G(\cdot)$, and introduce the RV's $\{M_K\}_1^\infty$ defined by

$$M_K := \max\{Y_1, ..., Y_K\}. \qquad K = 1, 2, ... \tag{7.11}$$

Also, for ease of notation, define the IR^+-valued sequence $\{m_K\}_1^\infty$ by

$$m_K := \inf\ \{\ x \geq 0 : 1 - G(x) \leq \frac{1}{K}\}. \qquad K = 1, 2, ... \tag{7.12}$$

The following result is a simplified version of Theorem 5(ii) given by Lai and Robbins [15, p. 103] adapted to the present set-up.

Theorem 7.3. *Let $\{Y_k\}_1^\infty$ be a family of i.i.d IR_+-valued RV's whose common distribution function $G(\cdot)$ satisfies the conditions*

$$G(x) < 1 \quad \text{for all } x \geq 0 \tag{7.13a}$$

and

$$\lim_{x \to +\infty} \frac{1 - G(cx)}{1 - G(x)} = 0 \quad \text{for all } c > 1. \tag{7.13b}$$

Under these conditions, the convergence

$$\lim_{K \to \infty} E\left[\left|\frac{M_K}{m_K} - 1\right|\right] = 0 \tag{7.14}$$

takes place, and the asymptotics

$$E[M_K] = m_K(1 + o(1)) \qquad K = 1, 2, ... \quad (7.15)$$

hold true with K going to infinity.

In view of Lemma 7.2, it is now natural to consider probability distribution functions $G(\cdot)$ with the tail behavior

$$P[Y_1 > x] = 1 - G(x) = Ce^{-qx}(1 + o(1)), \quad x \geq 0 \qquad (7.16)$$

for some $q > 0$ and $C > 0$. The next proposition summarizes the asymptotic properties associated with tail behavior (7.16). To simplify the exposition, for every $r \geq 1$, denote by $G_r(\cdot)$ the probability distribution function of the r-th power of any IR_+-valued RV distributed according to $G(\cdot)$, and in complete analogy with (7.12), define the real numbers $\{m_{K,r}\}_1^\infty$ by

$$m_{K,r} = \inf \{x \geq 0 : \ 1 - G_r(x) \leq \frac{1}{K}\}, \qquad K = 1, 2, ... \quad (7.17)$$

It is plain that

$$P[|Y_1|^r > x] = 1 - G_r(x) = Ce^{-qx^{\frac{1}{r}}}(1 + o(1)), \quad x \geq 0 \qquad (7.18)$$

with the identifications $G_r(\cdot) = G(\cdot)$ and $m_{K,r} = m_K$ taking place for $r = 1$. The main result of this section can now be given.

Theorem 7.4. *Let $\{Y_k\}_1^\infty$ be a family of i.i.d IR^+-valued RV's whose common distribution function $G(\cdot)$ exhibits the tail behavior (7.16). In that case, for all $r \geq 1$,*

(i): The probability distribution $G_r(\cdot)$ satisfies the conditions (7.13);

(ii): The asymptotic equivalence

$$\lim_{K \to \infty} \left[m_{K,r} \cdot [\frac{q}{\log K}]^r \right] = 1 \qquad (7.19)$$

holds true; and

(iii): The asymptotics

$$E\left[|M_K|^r\right] = [\frac{\log K}{q}]^r \cdot (1 + o(1)) \qquad (7.20)$$

hold true with K going to infinity.

Proof. Part (i) is readily checked by direct caculations which are omitted for sake of brevity. To show the equivalence (7.19), fix $r \geq 1$ and observe from (7.18) that for every $0 < \epsilon < 1$, there exists $x^*(\epsilon) > 0$ with the property that

$$1 - G_r^{-\epsilon}(x) \leq 1 - G_r(x) \leq 1 - G_r^{+\epsilon}(x) \qquad (7.21)$$

for all $x > x^*(\epsilon)$, with

$$G_r^{\pm\epsilon}(x) := 1 - Ce^{-qx^{\frac{1}{r}}}(1 \pm \epsilon), \quad x \geq 0. \qquad (7.22)$$

If the sequences $\{m_{K,r}^{\pm\epsilon}\}_1^\infty$ defined through (7.17) but with $G_r^{\pm\epsilon}(\cdot)$ instead of $G_r(\cdot)$, then simple computations show that

$$m_{K,r}^{\pm\epsilon} = [\frac{\log(K \cdot C(1 \pm \epsilon))}{q}]^r \qquad K = 1, 2, ... (7.23)$$

and the inequalities

$$m_{K,r}^{-\epsilon} \leq [\frac{\log(K \cdot C)}{q}]^r \leq m_{K,r}^{+\epsilon} \qquad K = 1, 2, ... (7.24)$$

are obtained. It also follows from (7.21) that for $K \geq K_\epsilon$ with some integer K_ϵ,

$$m_{K,r}^{-\epsilon} \leq m_{K,r} \leq m_{K,r}^{+\epsilon} \qquad (7.25)$$

and upon combining (7.24)-(7.25), the asymptotic equivalence (7.19) is obtained since obviously

$$\lim_{K \to \infty} \left[\frac{m_{K,r}^{+\epsilon}}{m_{K,r}^{-\epsilon}}\right] = 1 \qquad (7.26)$$

The asymptotics (7.15) of Theorem 7.3 and (7.19), when combined to the obvious relation

$$|M_K|^r = \max\{|Y_1|^r, \cdots, |Y_K|^r\}, \qquad (7.27)$$

readily yield (7.20). □

7.3. Asymptotics for the homogeneous FJ queue

A general asymptotic result for *homogeneous* FJ queues is now considered under the assumptions (R.1)-(R.8). and the synchronization constraint (2.10). Denote by (q, C) and $(\underline{q}, \underline{C})$ the constants defined in Lemma 7.2 for the pairs of Laplace-Stieltjes transforms $(A^*(\cdot), B^*(\cdot))$ and $(\underline{A}^*(\cdot), B^*(\cdot))$, respectively, where

$$\underline{A}^*(s) = \exp(-\frac{s}{\lambda}). \qquad (7.28)$$

The constants q and $\underline{q}$ are the smallest positive solutions to the equations

$$A^*(s)B^*(-s) = 1 \quad \text{and} \quad \underline{A}^*(s)B^*(-s) = 1, \tag{7.29}$$

respectively, and $q < \underline{q}$ by an easy convexity argument.

It follows from Lemma 7.2 that for all $1 \leq k \leq K$, the stationary response time R_∞^k in the k-th queue in the original FJ queue system has the tail

$$P[\ R_\infty^k > x\ ] = Ce^{-qx}(1 + o(1)), \quad x \geq 0 \tag{7.30}$$

whereas the stationary response time $\underline{R}_\infty^k$ of the k-th queue in the lower bound FJ queue system of Section 4.1 behaves like

$$P[\underline{R}_\infty^k > x] = \underline{C}e^{-\underline{q}x}(1 + o(1)), \quad x \geq 0. \tag{7.31}$$

These facts, when combined to the results of Section 7.2, provide insight into the growth behavior of the moments of the system response time, as the number of processors grows large. The main result along these lines is given in the following

Theorem 7.5. *Under the foregoing assumptions (R. 1)-(R.8), with the notation*

$$T_\infty^{(K)} := \max_{1 \leq k \leq K} R_\infty^k, \qquad K = 1, 2, \ldots \tag{7.32}$$

the bounds

$$\left[\frac{\log K}{\underline{q}}\right]^r \cdot (1 + o(1)) \leq E[|T_\infty^{(K)}|^r] \leq \left[\frac{\log K}{q}\right]^r \cdot (1 + o(1)) \tag{7.33a}$$

hold true for all $r \geq 1$, with K going to infinity, or equivalently

$$E[|T_\infty^{(K)}|^r] = O(\left[\log K\right]^r). \tag{7.33b}$$

This result generalizes a similar result obtained by Nelson and Tantawi [17] for the first moment of system response time in the case of exponential servers with Poissonian arrivals.

Proof. Now, with the terminology of Section 5, let $\{\overline{R}_\infty^1, \cdots, \overline{R}_\infty^K\}$ denote the *independent* version of the RV's $\{R_\infty^1, \cdots, R_\infty^K\}$. For notational convenience, pose

$$\underline{T}_\infty^{(K)} := \max_{1 \leq k \leq K} \underline{R}_\infty^k, \quad \overline{T}_\infty^{(K)} := \max_{1 \leq k \leq K} \overline{R}_\infty^k. \qquad K = 1, 2, \ldots \tag{7.34}$$

Note that $\underline{T}_\infty^{(K)} := T_\infty^2(\mathcal{D}_1)$ whereas $\overline{T}_\infty^{(K)} := T_\infty^{1,\{1,\cdots,K\}}$, and recall from the results of Sections 4 and 5 that $\underline{T}_\infty^{(K)} \leq_{cl} \overline{T}_\infty^{(K)}$ and $T_\infty^{(K)} \leq_{st} \overline{T}_\infty^{(K)}$, whence

$$E[|\underline{T}_\infty^{(K)}|^r] \leq E[|T_\infty^{(K)}|^r] \leq E[|\overline{T}_\infty^{(K)}|^r] \qquad K = 1, 2, ... (7.35)$$

since the mapping $IR_+ \to IR_+ : x \to x^r$ is monotone increasing and convex if $r \geq 1$.

As pointed out in Section 6, the RV's $\{|\underline{R}_\infty^1|^r, \cdots, |\underline{R}_\infty^K|^r\}$ are ***independent***. Owing to (7.31), Theorem 7.4 thus applies to yield

$$E[|\underline{T}_\infty|^r] = [\frac{\log K}{\underline{q}}]^r \cdot (1 + o(1)). \tag{7.36}$$

Similarly, the RV's $\{|\overline{R}_\infty^1|^r, \cdots, |\overline{R}_\infty^K|^r\}$ are ***independent***, and Theorem 7.4 again applies, this time by invoking (7.30), to give

$$E[|\overline{T}_\infty|^r] = [\frac{\log K}{q}]^r \cdot (1 + o(1)). \tag{7.37}$$

Combining (7.36) and (7.37) into (7.35) completes the proof. □

7.4 Homogeneous FJ queues with exponential servers

Assume the asumptions (R.1)-(R.8). With the notation (7.34), the expressions (6.23) and (6.30) now imply that

$$E[\underline{T}_\infty^{(K)}] = \frac{1}{\delta} H_K \quad \text{and} \quad E[\overline{T}_\infty^{(K)}] = \frac{1}{\theta} H_K \tag{7.38}$$

since $\pi = 1$ here by assumption (R.8). Moreover, $\delta = \mu(1 - \beta)$ and $\theta = \mu(1 - \alpha)$ with β and α denoting the smallest positive solution to the equations

$$b = \exp(-\frac{\mu(1-b)}{\lambda}), \quad b \geq 0, \tag{7.39}$$

and

$$a = A^*(\mu(1 - a)), \quad a \geq 0 \tag{7.40}$$

respectively. It is well known that

$$\log(K + 1) \leq H_K \leq 1 + \log K \qquad K = 1, 2, ... (7.41)$$

whence (7.33) thus obtains with $\underline{q} = \delta$ and $q = \theta$. This is to be expected as the reader can indeed check directly by a straightforward change of variable that when the probability distribution is exponential, the equations (7.28)-(7.29) reduce to (739)-(7.40).

APPENDIX A

A proof of Theorem 3.7

The proof proceeds in two steps which are defined by the value of the initial workload W. To indicate dependence on this initial workload, denote by $\{{}^{W}W_n\}_0^\infty$ the sequence of waiting times which are defined componentwise through the recursions (2.5) when the initial workload is W, i. e., for all $1 \leq k \leq K$,

$$ {}^{W}W_{n+1}^k = [{}^{W}W_n^k + u_n^k \cdot \sigma_n^k - \tau_{n+1}^k]^+ \qquad n = 0, 1, ...(A.1) $$

with ${}^{W}W_0 = W$.

The first step is the one where the initial workload W is zero, in which case iterating (A.1) yields the well-known representation

$$ W_n^k = \max\{0, O_n^k, O_n^k + O_{n-1}^k, \cdots, O_n^k + ... + O_1\} \qquad n = 1, 2, ...(A.2) $$

with the RV's $\{O_n^k\}_0^\infty$ as defined in (3.18).

Following Loynes [16], it is convenient to embed the sequence $\{O_n\}_1^\infty$ into a larger *stationary ergodic* sequence, say $\{O_n\}_{-\infty}^{+\infty}$. Such an extension is possible owing to the enforced assumption (A.3). The IR_+^K-valued RV's $\{V_n\}_0^\infty$ are now defined componentwise by

$$ V_n^k := \max\{0, O_{-1}^k, O_{-1}^k + O_{-2}^k, \cdots, O_{-1}^k + \cdots + O_{-n}^k\}, \qquad n = 1, 2, ...(A.3) $$

with $V_0 = 0$, and equivalence in law being denoted by $=_{st}$, it is clear that

$$ {}^{0}W_n =_{st} V_n. \qquad n = 0, 1, ...(A.4) $$

The ergodicity assumption now entails

$$ \lim_{n \to -\infty} \frac{1}{n} \sum_{i=1}^{n} O_{-i} = E[O_0] < 0 \quad a.s. \qquad (A.5) $$

where the last inequality follows from the stability condition (3.42). Consequently, the convergence

$$ \lim_{n \to -\infty} \sum_{i=1}^{n} O_{-i}^k = -\infty \quad a.s. \qquad (A.6) $$

takes place for all $1 \leq k \leq K$, and implies the existence of an a.s. *finite* random rank N_k with the property that for all $n > N_k$,

$$\sum_{i=1}^{n} O_{-i}^{k} < 0 \quad a.s. \tag{A.7}$$

The reader will easily check from the defining relation (A.3) that the sequence of RV's $\{V_n\}_1^\infty$ is componentwise ***monotone non-decreasing***, i. e., for all $1 \leq k \leq K$,

$$0 \leq V_{n-1}^{k} \leq V_n^{k} \qquad n = 1, 2, ... \tag{A.8}$$

and the RV V_∞ whose k-th component is given by

$$V_\infty^{k} := \lim_{n \to -\infty} V_n^{k}, \quad 1 \leq k \leq K, \tag{A.9}$$

is thus well defined. Owing to (A.7), this RV V_∞^k is readily interpreted as the maximum of an a.s. *finite* number of RV's, whence it is a.s. finite, and the sequence of IR_+^K-valued RV's $\{V_n\}_0^\infty$ converges a.s. to an a.s. *finite* RV V_∞, whose k-th component is given by (A.9). Consequently, owing to (A.4), the IR_+^K-valued RV's $\{{}^0W_n\}_0^\infty$ necessarily converge ***weakly*** to an a.s. finite IR_+^K-valued RV, say W_∞, which is identical in law with the non-defective RV V_∞.

For the case of a non-zero initial workload $W \geq 0$, an easy induction argument shows that

$${}^W W_n^k \geq {}^0 W_n^k, \qquad n = 0, 1, ... \tag{A.10}$$

and the RV ν_k given by

$$\nu_k := \inf\{n \geq 0 : {}^W W_n^k = 0\}, \tag{A.11}$$

is thus a.s. *finite* under the stability condition (3.42). Therefore, ${}^W W_n^k = {}^0 W_n^k$ necessarily for all $n \geq \nu_k$ by virtue of (A.1) and (A.10) with ${}^0 W_{\nu_k}^k = 0$. The RV ν given by

$$\nu = \max_{1 \leq k \leq K} \nu_k \tag{A.12}$$

is thus a.s. *finite* and has the property that ${}^W W_n = {}^0 W_n$ for all $n \geq \nu$.

The first part of the proof now implies that the sequence of IR_+^K-valued RV's $\{{}^W W_n\}_0^\infty$ converges ***weakly*** to the a.s. finite (and thus non-defective) RV W_∞, i. e., Lemma 3.7. is indeed obtained for the sequence $\{W_n\}_0^\infty$. It is clear from the discussion that the limiting distribution is independent of the initial workload distribution. The corresponding result for the sequence of

IR_+^K-valued RV's $\{W_n(\mathcal{D})\}_0^\infty$ is obtained by similar arguments which are omitted for sake of brevity. □

APPENDIX B

A proof of Theorem 3.8

Let ϕ be any convex non-decreasing function $\phi : IR^K \to IR^+$ with the property that

$$E[\phi(W_\infty)] < \infty \text{ and } E[\phi(W_\infty(\mathcal{D}))] < \infty. \tag{B.1}$$

The monotonicity property (A.8) readily implies that the RV's $\{V_n\}_0^\infty$ defined by the pathwise scheme (A.2), satisfy the inequalities

$$V_n^k \leq V_{n+1}^k \leq V_\infty^k \qquad n = 0, 1, ... (B.2)$$

for all $1 \leq k \leq K$. The Monotone Convergence Theorem thus yields

$$\begin{aligned} E[\phi({}^0W_n)] &= E[\phi(V_n)] \\ &\leq \lim_{n\to\infty} E[\phi(V_n)] = E[\phi(V_\infty)] = E[\phi(W_\infty)] \end{aligned} \qquad n = 0, 1, ... (B.3)$$

upon making use of the results of Appendix A. A similar result is obtained for the sequence $\{{}^0W_n(\mathcal{D})\}_0^\infty$, provided it is stationary, i.e., with obvious notation,

$$\begin{aligned} E[\phi({}^0W_n(\mathcal{D}))] &= E[\phi(V_n(\mathcal{D}))] \\ &\leq \lim_{n\to\infty} E[\phi(V_n\mathcal{D})] = E[\phi(V_\infty(\mathcal{D}))] = E[\phi(W_\infty(\mathcal{D}))] \end{aligned} \qquad n = 0, 1, ... (B.4)$$

The RV's $\phi(V_n)$ and $\phi(V_n(\mathcal{D}))$ are thus *integrable* as a result of (B.1). Furthermore, Corollary 3.5 and (A.4) give the relation

$$E[\phi(V_n(\mathcal{D}))] \leq E[\phi(V_n)]. \qquad n = 0, 1, ... (B.5)$$

The proof of (3.44) is now completed by letting n go to infinity in (B.5) and by making use of (B.3) and (B.4). The inequalities (3.45)-(3.47) are now immediate under the enforced assumptions. □

APPENDIX C

Single server queues with Bernoulli loading

This appendix is devoted to the derivation of some useful facts for single server queues with Bernoulli loading sequences. To define such queueing systems, let the three sequences of non-negative *integrable* RV's $\{\sigma_n\}_0^\infty$, $\{\tau_n\}_1^\infty$ and $\{u_n\}_0^\infty$ be *mutually independent* sequences of *i.i.d.*

RV's, which are assumed *independent* of an additional IR_+-valued RV W. Moreover, the RV's $\{u_n\}_0^\infty$ form a *Bernoulli* sequence with

$$P[u_n = 1] = 1 - P[u_n = 0] = \pi \qquad n = 0, 1, ...(C.1)$$

where $0 < \pi \leq 1$. Also let $A(\cdot)$ and $B(\cdot)$ be the common probability distribution functions of the sequences of interarrival times $\{\tau_m\}_1^\infty$ and service durations $\{\sigma_m\}_0^\infty$, respectively, and denote their first moment by $\overline{\tau}$ and $\overline{\sigma}$, respectively.

The corresponding GI/GI/1 system with Bernoulli loading sequence is the single server system defined by the interarrival times $\{\tau_n\}_1^\infty$ and service requirements $\{u_n \cdot \sigma_n\}_0^\infty$. Its sequence $\{W_n\}_0^\infty$ of waiting times is generated by the recursion

$$W_{n+1} = [W_n + u_n \cdot \sigma_n - \tau_{n+1}]^+ \qquad n = 0, 1, ...(C.2)$$

with $W_0 = W$.

A general result

Another single server queueing system can be obtained from the original system by changing the original time scale (which is defined by tagging incoming customers) to account only for the times at which customers indeed do join the queue. This is done by introducing the RV's $\{k_m\}_0^\infty$ which are recursively defined by

$$k_0 := \inf\{k \geq 0 : u_k = 1\} \qquad (C.3)$$

and

$$k_{m+1} := \inf\{k > k_m : u_k = 1\} \qquad m = 0, 1, ...(C.4)$$

with the usual convention that when the defining set in (C.3)-(C.4) is empty, the RV is set equal to ∞. It is clear that the RV's $\{k_m\}_0^\infty$ are finite a.s.

The IR_+-valued RV $\{\tilde{\sigma}_m\}_0^\infty$ and $\{\tilde{\tau}_m\}_1^\infty$ are now defined by

$$\tilde{\sigma}_m = \sigma_{k_m} \qquad m = 0, 1, ...(C.5)$$

and

$$\tilde{\tau}_{m+1} = \sum_{k_m \leq \ell < k_{m+1}} \tau_{\ell+1}. \qquad m = 0, 1, ...(C.6)$$

The auxiliary GI/GI/1 queue is the one associated with the sequence of interarrival times $\{\tilde{\tau}_m\}_1^\infty$ and service requirements $\{\tilde{\sigma}_m\}_0^\infty$, and its sequence of waiting times $\{\tilde{W}_m\}_0^\infty$ is now defined through the recursion

$$\tilde{W}_{m+1} = [W_m + \tilde{\sigma}_m - \tilde{\tau}_{m+1}]^+ \qquad m = 0, 1, ...(C.7)$$

with $\tilde{W}_0 = W_{k_0}$.

A moment of reflection should convince the reader that the second system is obtained from the original upon sampling it along the time sequence $\{k_m\}_0^\infty$, since $\tilde{W}_m = W_{k_m}$ for all $m = 0, 1, ...$ as readily checked from the Lindley recursions (C.2) and (C.7). The following lemma further explores the probabilistic implications of this relationship.

Lemma C.1. *Both* GI/GI/1 *systems are stable under the condition* $\pi\overline{\sigma} < \overline{\tau}$, *in the sense that the sequences of waiting times* $\{W_n\}_0^\infty$ *and* $\{\tilde{W}_m\}_0^\infty$ *converge weakly to some non-defective probability distributions on* IR_+, *say* $F(\cdot)$ *and* $\tilde{F}(\cdot)$, $F(\cdot)$ *and* $\tilde{F}(\cdot)$ *are independent of the initial waiting time distribution and coincide, i. e.,* $F(\cdot) = \tilde{F}(\cdot)$.

Proof. The RV's $\{\nu_m\}_1^\infty$ defined by

$$\nu_{m+1} := \begin{cases} k_{m+1} - k_m & \text{if } k_m < \infty \\ +\infty & \text{otherwise} \end{cases} \qquad m = 0, 1, ...(C.8)$$

form a sequence of i.i.d. *geometric* RV's which are independent of the RV k_0. In fact, a simple argument shows that

$$\begin{aligned} P[\nu_{m+1} = j + 1] &= P[u_{k_m+i} = 0, 1 \le i \le j, u_{k_m+j+1} = 1] \\ &= \pi(1 - \pi)^j \end{aligned} \qquad (C.9)$$

for all $j = 0, 1, ...$ and the relation

$$E[\nu_{m+1}] = \sum_{j=0}^{\infty}(j + 1)\pi(1 - \pi)^j = \frac{1}{\pi} \qquad (C.10)$$

obtains by elementary computations.

Given the enforced independence assumptions, the relations $E[u_n \cdot \sigma_n] = \pi\overline{\sigma}$, $E[\tilde{\sigma}_m] = \overline{\sigma}$ and $E[\tilde{\tau}_m] = \frac{1}{\pi}\overline{\tau}$ are easily checked to hold for all $n, m = 0, 1, ...$, with the last one following by a direct application of Wald's identity. For both systems, the well-known stability condition for GI/GI/1 queues thus simply reduces to the stated inequality $\pi\overline{\sigma} < \overline{\tau}$. It is well known [10] that under

this stability condition, the sequences of RV's $\{W_n\}_0^\infty$ and $\{\tilde{W}_m\}_0^\infty$ converge weakly to some non-defective probability distributions on IR_+, say $F(\cdot)$ and $\tilde{F}(\cdot)$, which are independent of the initial waiting time probability distribution function.

The strong Ergodic Theorem readily implies that the convergence

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} e^{-sW_k} = F^*(s) \quad a.s. \tag{C.11}$$

and

$$\lim_{m\to\infty} \frac{1}{m+1} \sum_{\ell=0}^{m} e^{-s\tilde{W}_\ell} = \tilde{F}^*(s) \quad a.s. \tag{C.12}$$

for all $s \geq 0$.

On the other hand, it is clear that

$$\begin{aligned} \frac{1}{m+1} \sum_{\ell=0}^{m} e^{-s\tilde{W}_\ell} &= \frac{1}{m+1} \sum_{\ell=0}^{m} e^{-sW_{k_\ell}} \\ &= \frac{1}{m+1} \sum_{n=0}^{k_m} u_n e^{-sW_n} \qquad m = 0,1,... \quad (C.13) \\ &= \left(\frac{k_m+1}{m+1}\right) \cdot \frac{1}{k_m+1} \sum_{n=0}^{k_m} u_n e^{-sW_n}, \end{aligned}$$

whereas the obvious identity

$$k_m = k_0 + \nu_1 + ... + \nu_m \qquad m = 0,1,... \quad (C.14)$$

and the Strong Law of Large Numbers now immediately yield

$$\lim_{m\uparrow\infty} \frac{k_m+1}{m+1} = E[\nu_1] = \frac{1}{\pi} \quad a.s. \tag{C.15}$$

The RV's W_n and u_n are clearly independent for all $n = 0,1,...$ and the Strong Ergodic Theorem again applies to give

$$\lim_{N\uparrow\infty} \frac{1}{N+1} \sum_{n=0}^{N} u_n e^{-sW_n} = \pi F^*(s) \quad a.s. \tag{C.16}$$

Since

$$k_m + 1 \leq k_{m+1} < \infty \quad a.s. \qquad m = 0,1,... \quad (C.17)$$

it is plain that $\lim_{m\uparrow\infty} k_m = \infty$ a.s., and standard arguments applied to (C.16) give

$$\lim_{m\uparrow\infty} \frac{1}{k_m+1} \sum_{n=0}^{k_m} u_n e^{-sW_n} = \pi F^*(s) \quad a.s. \tag{C.18}$$

Consequently, upon combining (C.13), (C.15) and (C.18), the reader will check that

$$\lim_{m\uparrow\infty} \frac{1}{m+1} \sum_{\ell=0}^{m} e^{-s\tilde{W}_\ell} = \frac{1}{\pi} \cdot \pi F^*(s) \quad a.s. \tag{C.19}$$

and the equality

$$\tilde{F}^*(s) = F^*(s) \tag{C.20}$$

is obtained for all $s \geq 0$ by direct comparison with (C.12), i. e., $F(\cdot) = \tilde{F}(\cdot)$. □

Special cases

Although the *service durations* $\{\sigma_m\}_0^\infty$ are distributed according to the distribution $B(\cdot)$, the i.i.d *effective service durations* (or service requirements) $\{u_n \cdot \sigma_n\}_0^\infty$ are all distributed according to the related distribution $\tilde{B}(\cdot)$ given by

$$\tilde{B}(x) = \pi B(x) + (1-\pi)B(0+), \quad x \geq 0 \tag{C.22}$$

Thus, in some sense, the single server queue with Bernoulli loading is nothing more than a GI/GI/1 queue with interarrival time and service duration distributions given by $A(\cdot)$ and $\tilde{B}(\cdot)$, respectively. However, the usefulness of Lemma C.1 lies in showing that the computation of the equilibrium waiting distribution for the single server queueing system with Bernoulli loading is *equivalent* to the computation of the corresponding distribution for a *standard* GI/GI/1 system with interarrival and service time distributions $\tilde{A}(\cdot)$ and $B(\cdot)$, respectively, with the Laplace-Stieltjes transform $\tilde{A}^*(\cdot)$ of the former given by

$$\begin{aligned} \tilde{A}^*(s) &= \sum_{j=0}^{\infty} \pi(1-\pi)^j A^{*(j+1)}(s) \\ &= \frac{\pi A^*(s)}{1-(1-\pi)A^*(s)} \end{aligned} \tag{C.22}$$

for all $s \geq 0$. This follows by a standard argument based on the enforced independence together with (C.6) and (C.10).

Two special cases are now considered. If the arrivals to the system are modelled by a *Poisson* process, say of rate λ, then

$$A^*(s) = \frac{\lambda}{\lambda + s} \quad \text{and} \quad \tilde{A}^*(s) = \frac{\pi\lambda}{\pi\lambda + s} \tag{C.23}$$

for all $s \geq 0$. Thus, as expected, the equilibrium waiting time distribution in a single server queueing system with Bernoulli loading, Poissonan input (of rate λ) and service time distribution $B(\cdot)$ is exactly the equilibrium waiting time distribution in a standard M/GI/1 queue with arrival rate $\pi\lambda$ and service time distribution $B(\cdot)$.

The second situation of interest occurs when the distribution $B(\cdot)$ is an *exponential* distribution, say with parameter μ, in which case evaluation of the probability distribution $F(\cdot)$ amounts to solving for the equilibrium waiting time distribution in an auxiliary GI/M/1 system. In fact, well-known results on GI/M/1 systems [10] imply that

$$F(t) = 1 - \alpha e^{-\theta t}, \quad t \geq 0 \tag{C.24}$$

where α is the smallest positive solution to the equation

$$x = \tilde{A}^*(\mu(1 - x)), \quad x \geq 0 \tag{C.25a}$$

and

$$\theta := \mu(1 - \alpha). \tag{C.25b}$$

The corresponding Laplace-Stieltjes transform $F^*(\cdot)$ is then given by

$$F^*(s) = (1 - \alpha) + \alpha \cdot \frac{\theta}{\theta + s} \tag{C.26}$$

for all $s \geq 0$.

REFERENCES

[1] F. Baccelli, *Two parallel queues created by arrivals with two demands*, Rapport de Recherche No. 426, INRIA - Rocquencourt (France), 1985.

[2] F. Baccelli and A.M. Makowski, "Simple computable bounds for the Fork-Join queue," *Proceedings of the 19th Annual Conference on Information Sciences and Systems*, The Johns Hopkins University, Baltimore, Maryland, March 1985, pp. 436-441.

[3] F. Baccelli and A.M. Makowski, "Asymptotics for the Fork-Join queue", Abstract, Workshop on Computer Performance Evaluation, Sophia-Antipolis, April 1986.

[4] F. Baccelli and A.M. Makowski, "Properties of stochastic ordering for associated random variable", *Opns. Res.*, under revision, 1986.

[5] F. Baccelli, A.M. Makowski and A. Shwartz, "Simple Computable bounds and approximations for the Fork-Join queue", *Proceedings of the International Workshop on Computer Performance Evaluation*, Tokyo, Japan, September 1985, pp. 437-450.

[6] F. Baccelli and A.M. Makowski, "Stability and bounds for single server queue in random environment", *Stochastic Models* vol. 2 (1986), pp. 281-292.

[7] F. Baccelli and W.A. Massey, *Series- parallel Fork-Join queueing networks and their stochastic ordering*, Rapport de Recherche No. 488, INRIA - Rocquencourt (France), 1986.

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

[9] A.A. Borovkov, *Stochastic Processes in Queueing Theory*, English Translation, Springer-Verlag, New York - Berlin, 1976.

[10] J.W. Cohen, *The Single Server Queue*, North-Holland, Amsterdam, 1969.

[11] J.D. Esary, F. Proschan and D.W. Walkup, "Association of random variables, with applications", *Ann. Math. Stat.* vol. 38 (1967), pp. 1466-1474.

[12] L. Flatto and S. Hahn, "Two parallel queues created by arrivals with two demands I," *SIAM J. Appl. Math.* vol 44 (1984), pp. 1041-1053.

[13] B. Hajek, "The proof of a folk theorem on queueing delay with applications to routing in networks," *J. Assoc. Comp. Mach.* vol. 30 (1883), pp. 834-851.

[14] P. Humblet, *Determinism minimizes waiting times in queues*, Technical Report, LIDS - Department of Electrical Engineering and Computer Science, MIT, 1982.

[15] T.L. Lai and H. Robbins, "A class of dependent random variables and their maxima," *Z. Wahr. verw. Geb.* vol. 42 (1978), pp. 89-111.

[16] R.M. Loynes, "The stability of a queue with non-independent interarrival and service times," *Proc. Cambridge Philo. Soc.* vol. 5 (1962), pp. 497-520.

[17] R. Nelson and A. Tantawi, IBM Research Report, Yorktown Heights, New York, 1985.

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

[19] S. Ross, *Stochastic Processes*, J. Wiley & Sons, New York, 1984.

[20] T. Rolski, *Comparison theorems fo queues with dependent inter-arrival times*, Modelling and Performance Evaluation Methodology, Lecture Notes in Control and Information Sciences No 60, Springer Verlag, 1984.

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

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

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique