Copyright © 1981 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 60, No. 1, January 1981
Printed in U.S.A.

Resource Sharing for Efficiency in Traffic Systems

By D. R. SMITH and W. WHITT

(Manuscript received June 27, 1980)

Experience has shown that efficiency usually increases when separate traffic systems are combined into a single system. For example, if Group A contains 10 trunks and Group B 8 trunks, there should be fewer blocked calls if A and B are combined into a single group of 18 trunks. It is intuitively clear that the separate systems are less efficient because a call can be blocked in one when trunks are idle in the other. Teletraffic engineers and queuing theorists widely accept such efficiency principles and often assume that their mathematical proofs are either trivial or already in the literature. This is not the case for two fundamental problems that concern combining blocking systems (as in the example above) and combining delay systems. For the simplest models, each problem reduces to the proof of an inequality involving the corresponding classical Erlang function. Here the two inequalities are proved in two different ways by exploiting general stochastic comparison concepts: first, by monotone likelihood-ratio methods and, second, by sample-path or "coupling" methods. These methods not only yield the desired inequalities and stronger comparisons for the simplest models, but also apply to general arrival processes and general service-time distributions. However, it is assumed that the service-time distributions are the same in the systems being combined. This common-distribution condition is crucial since it may be disadvantageous to combine systems with different service-time distributions. For instance, the adverse effect of infrequent long calls in one system on frequent short calls in the other system can outweigh the benefits of making the two groups of servers mutually accessible.

I. INTRODUCTION AND SUMMARY

From extensive experience in teletraffic engineering, it is well known that congestion can often be reduced by sharing resources. The blocking probability in a loss system and the average waiting time in a delay system are usually much less when separate facilities serving separate streams of traffic are combined to serve all the streams together. Alternatively, for a given level of congestion, fewer facilities are usually required to serve the streams together. Sometimes such results are trivial: Whenever the combined system may be managed as if it were in fact separate systems, the optimal performance of the combined system is at least as good as that of the separate systems. However, such management is not allowed in the models treated here. In any case, the efficiency of shared resources is certainly a fundamental principle of teletraffic engineering.

The purpose of this paper is to establish versions of this efficiency principle mathematically. Our first two results verify conjectures by Arthurs and Stuck.[1] To state our first result, let $L(s, \lambda\ \mu)$ denote the stationary loss or overflow rate in an $M/M/s$ loss system (no waiting room) with s servers, arrival rate λ, and individual service rate μ. (See Kleinrock[2] for background on the queuing models.) It is well known that $L(s, \lambda, \mu) = \lambda B(s, a)$, where $a = \lambda/\mu$ and $B(s, a)$ is the familiar Erlang blocking formula:

$$B(s, a) = (a^s/s!)/\sum_{k=0}^{s} (a^k/k!); \tag{1}$$

see Jagerman[3] and references there. The first efficiency principle we establish says that $L(s, \lambda, \mu)$ is a subadditive function of (s, λ) for each fixed μ:

Theorem 1: For all positive integers s_1 and s_2 and all positive real numbers λ_1, λ_2 and μ,

$$L(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \le L(s_1, \lambda_1, \mu) + L(s_2, \lambda_2, \mu). \tag{2}$$

This yields immediately that

$$G\left(\sum_{i=1}^{n} s_i, \sum_{i=1}^{n} a_i\right) \le \sum_{i=1}^{n} G(s_i, a_i)$$

for each integer n, where $G(s, a) = aB(s, a)$, which is the version of Theorem 1 actually conjectured by Arthurs and Stuck.[1]

Of course, Theorem 1 should not surprise teletraffic engineers, since it can be inferred from common tables and graphs, but it has apparently not been proved before. It appears that all previous mathematical results can be described as "one-parameter" results. The relation (2) has been deduced for special cases in which quantities such as the blocking probability or the load per server are held constant. For example, it is known that if one combines separate groups of j and k trunks, each operating at a blocking probability 0.01, then the new blocking probability will be less than 0.01 (or, alternatively, the combined system can handle an increased total load and retain the same

0.01 blocking probability); see p. 68 of Cooper.[4] Such results are often presented without rigorous mathematical support.

From Paul Burke we learned about another special case that has been known for a long time. It is not difficult to show that $B(ts, ta)$ is strictly decreasing in t (see the appendix), from which (2) easily follows in the case $\lambda_1/s_1 = \lambda_2/s_2$. Herbert Shulman has also shown that Theorem 1 follows easily from the monotonicity of $B(ts, ta)$ in t and the convexity of $B(s, a)$ in s for $s \geq 1$, but such convexity has not yet been established (see the appendix). For further discussion of other related work, see Section 5.1 of Kleinrock.[5]

To state our second result, let $D(s, \lambda, \mu)$ denote the mean steady-state delay in an $M/M/s$ queue with infinite waiting room, FCFS (first-come, first-served) queue discipline, s servers, arrival rate λ, and individual service rate μ. It is well known that $D(s, \lambda, \mu) = C(s, \lambda/\mu)/(s\mu - \lambda)$, where $C(s, a)$ is the Erlang delay function:

$$C(s, a) = \frac{a^s/(s-1)!\,(s-a)}{\sum_{k=0}^{s-1} (a^k/k!) + a^s/(s-1)!\,(s-a)}. \tag{3}$$

The following result establishes subadditivity of $D(s, \lambda, \mu)$ as a function of (s, λ) for each fixed μ. Note that

$$[\lambda_1/(\lambda_1 + \lambda_2)]D(s_1, \lambda_1, \mu) + [\lambda_2/(\lambda_1 + \lambda_2)]D(s_2, \lambda_2, \mu)$$

is the overall average delay experienced in the separate systems because $\lambda_1/(\lambda_1 + \lambda_2)$ is the long-run proportion of customers to enter the first system.

Theorem 2: For all positive integers s_1 and s_2 and all positive real numbers λ_1, λ_2, and μ,

$$\begin{aligned} &D(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \\ &\qquad \leq [\lambda_1/(\lambda_1 + \lambda_2)]D(s_1, \lambda_1, \mu) + [\lambda_2/(\lambda_1 + \lambda_2)]D(s_2, \lambda_2, \mu). \end{aligned} \tag{4}$$

This yields immediately that

$$H\left(\sum_{i=1}^{n} s_i, \sum_{i=1}^{n} a_i\right) \leq \sum_{i=1}^{n} H(s_i, a_i)$$

for each integer n, where $H(s, a) = aC(s, a)/(s - a)$, which is the version of Theorem 2 actually conjectured by Arthurs and Stuck.[1]

In order to prove Theorems 1 and 2, we found it convenient to prove stronger results. It is helpful to see how the loss rate $L(s, \lambda, \mu)$ and the mean delay $D(s, \lambda, \mu)$ are related to the steady-state number of customers in the system, say Q. In the $M/M/s$ loss system

$$L(s, \lambda, \mu) = \lambda - \mu EQ \tag{5}$$

because λ is the arrival rate and μEQ is the service-completion rate;

the loss rate is that fraction of the arrivals that are not served. In the $M/M/s$ delay system

$$\lambda(D(s, \lambda, \mu) + \mu^{-1}) = EQ \tag{6}$$

by virtue of the fundamental relation $L = \lambda W$; see Stidham.[6]

Let Q_i be the steady-state number of customers in the ith system ($i = 1, 2$) and let Q be the steady-state number of customers in the combined system. Then Theorem 1 is equivalent to

$$EQ \geq EQ_1 + EQ_2 \tag{7}$$

for the loss systems, and Theorem 2 is equivalent to

$$EQ \leq EQ_1 + EQ_2 \tag{8}$$

for the delay systems.

Instead of comparing the means in (7) and (8), we prove Theorems 1 and 2 by making more general stochastic comparisons. We do this in two different ways. Our first method of proof is to compare the distribution of Q with the distribution of $Q_1 + Q_2$. It turns out to be very easy to establish an appropriate ordering for the entire distributions, which in turn implies the desired inequality for the means. The appropriate order is the monotone likelihood-ratio ordering. We define this ordering and prove the more general theorems implying Theorems 1 and 2 in Section II.

Our second method of proof is to compare entire stochastic processes rather than just stationary distributions. As corollaries we obtain stochastic-order relations for the stationary distributions which in turn also imply the desired inequalities (7) and (8) for the means. This approach has the advantage that the arrival processes can be arbitrary rather than Poisson and the service-time distributions can be general instead of exponential. The argument is also remarkably simple. The idea in this approach is to construct artificially the two stochastic processes being compared on the same probability space. The construction is carried out so that each stochastic process individually has the correct distribution (family of finite-dimensional distributions) as originally specified. We choose a special joint distribution so that each sample path of one process always lies below the corresponding sample path of the other process. Because the construction is artificial, the joint distribution of the two processes is not directly meaningful, but it implies a strong stochastic ordering for the processes. Such special constructions have been used previously to compare queuing processes; see Sonderman,[7] Whitt,[8] Wolff,[9] and references there. In fact, the generalization of Theorem 2 is a direct consequence of Wolff's theorem and the other proofs involve similar reasoning. We present our results using this approach in Section III.

In Theorems 1 and 2 we assume equal service rates in the two systems. It is natural to ask whether extensions of (2) and (4) hold when the service rates are unequal. In Section IV we show that, with unequal service rates, combining resources need not be more efficient; in fact it can substantially degrade performance. Infrequent "bad" customers from one system can adversely affect a large number of "good" customers from the other system.

II. MONOTONE LIKELIHOOD-RATIO COMPARISONS

Let X and Y be random variables assuming values in the nonnegative integers. We say X is less than or equal to Y in the monotone likelihood-ratio ordering and write $X \leq_r Y$ if

$$\frac{P(X = k + 1)}{P(X = k)} \leq \frac{P(Y = k + 1)}{P(Y = k)} \tag{9}$$

for all integers k; see page 208 of Ferguson.[10] We say X is stochastically less than or equal to Y and write $X \leq_{st} Y$ if $Ef(X) \leq Ef(Y)$ for all nondecreasing real-valued functions f for which the expectations are well defined. Obviously, $EX \leq EY$ whenever $X \leq_{st} Y$. What is important for Theorems 1 and 2 is that $X \leq_r Y$ implies $X \leq_{st} Y$. This is well known and not difficult to show. In fact, the monotone likelihood-ratio ordering is equivalent to stochastic order for all conditional distributions obtained by conditioning on subsets, i.e., $E(f(X) \mid X \in A) \leq E(f(Y) \mid Y \in A)$ for all subsets A and all nondecreasing real-valued functions f; this property is discussed in Whitt[11,12]; see Keilsoñ and Sumita[13] for additional material.

Returning to the notation of (7) and (8), we obtain the following results which imply Theorems 1 and 2.

Theorem 3: For the M/M/s loss systems,

$$Q_1 + Q_2 <_r Q.$$

Theorem 4: For the M/M/s delay systems,

$$Q <_r Q_1 + Q_2.$$

Theorems 3 and 4 can each be proved by simple calculations since the stationary distributions are known and easy to work with. To illustrate, we do one proof.

Direct Proof of Theorem 3: Let $a_i = \lambda_i/\mu_i$ for $i = 1, 2$. Then, using convolution, we obtain for some constant C

$$P(Q_1 + Q_2 = k + 1) = C \sum_{\substack{0 \leq i_1 \leq s_1 \\ 0 \leq i_2 \leq s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1} a_2^{i_2} / i_1! i_2!$$

$$= \frac{Ca_1}{k+1} \sum_{\substack{1 \le i_1 \le s_1 \\ 0 \le i_2 \le s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1 - 1} a_2^{i_2} / (i_1 - 1)! i_2!$$

$$+ \frac{Ca_2}{k+1} \sum_{\substack{0 \le i_1 \le s_1 \\ 1 \le i_2 \le s_2 \\ i_1 + i_2 = k+1}} a_1^{i_1} a_2^{i_2 - 1} / i_1! (i_2 - 1)!$$

$$< \left(\frac{a_1 + a_2}{k+1} \right) C \sum_{\substack{0 \le i_1 \le s_1 \\ 0 \le i_2 \le s_2 \\ i_1 + i_2 = k}} a_1^{i_1} a_2^{i_2} / i_1! i_2!$$

$$= \frac{P(Q = k+1)}{P(Q = k)} P(Q_1 + Q_2 = k). \qquad \square$$

It is also significant that both Theorems 3 and 4 can be viewed as trivial corollaries of a more general theorem. This more general theorem is especially useful for comparisons when the limiting distributions are not known. To state our general result, consider two stochastic processes on the integers, $Y_1(t)$ and $Y_2(t)$, that move only by jumps up or down in unit steps to one of the neighboring states. Let all the transitions be governed by birth-and-death rates, but in contrast to those in birth-and-death processes, these rates may depend on information other than the current state such as the history of the process or other relevant variables. Let $\lambda_i(k, I_t)$ and $\mu_i(k, I_t)$ be the birth-and-death rates, respectively, for the ith process ($i = 1, 2$) in state k with additional information I_t at time t. By having transitions governed by birth-and-death rates, we mean that

$$P(Y_i(t+h) = k+1 \mid Y_i(t) = k, I_t) = h\lambda_i(t, I_t) + o(h),$$
$$P(Y_i(t+h) = k-1 \mid Y_i(t) = k, I_t) = h\mu_i(t, I_t) + o(h),$$

and

$$P(Y_i(t+h) = k \mid Y_i(t) = k, I_t) = 1 - h[\lambda_i(t, I_t) + \mu_i(t, I_t)] + o(h),$$

where $o(h)$ means a quantity that converges to zero after division by h as $h \to 0$. Let X_1 and X_2 be random variables with the limiting distributions of these two stochastic processes, which we assume exist as proper distributions. Here is our general monotone likelihood-ratio comparison result.

Theorem 5: Consider the processes $Y_1(t)$ and $Y_2(t)$ defined above. Suppose there exist sequences of constants $\{\alpha_i(k))\}$ and $\{\beta_i(k))\}$ such that

$$\lambda_1(k, I_t) \le \alpha_1(k), \qquad \lambda_2(k, I_t) \ge \alpha_2(k),$$
$$\mu_1(k, I_t) \ge \beta_1(k), \qquad \mu_2(k, I_t) \le \beta_2(k),$$

for all k and I_t. If $\alpha_1(k)/\beta_1(k+1) \leq \alpha_2(k)/\beta_2(k+1)$ for all k, then $X_1 \leq_r X_2$.

Corollary: If

$$\lambda_1(k, I_t) \leq \lambda_2(k, I'_t)$$

and

$$\mu_1(k, I_t) \geq \mu_2(k, I'_t)$$

for all k, I_t, and I'_t, then $X_1 \leq_r X_2$.

Proof of Theorem 5: Look at the stationary flow between states k and $k+1$. The flow from k to $k+1$ is less than or equal to $P(X_1 = k)\alpha_1(k)$ for process 1 and greater than or equal to $P(X_2 = k)\alpha_2(k)$ for process 2. Similarly, the stationary flow from $k+1$ to k is greater than or equal to $P(X_1 = k+1)\beta_1(k)$ in process 1 and less than or equal to $P(X_2 = k+1)\beta_2(k)$ in process 2. Since the stationary flow from k to $k+1$ must equal the stationary flow in the opposite direction,

$$P(X_1 = k)\alpha_1(k) \geq P(X_1 = k+1)\beta_1(k+1)$$

and

$$P(X_2 = k)\alpha_2(k) \leq P(X_2 = k+1)\beta_2(k+1).$$

Consequently,

$$\frac{P(X_1 = k+1)}{P(X_1 = k)} \leq \frac{\alpha_1(k)}{\beta_1(k+1)} \leq \frac{\alpha_2(k)}{\beta_2(k+1)} \leq \frac{P(X_2 = k+1)}{P(X_2 = k)}. \qquad \square$$

We can now apply the corollary to Theorem 5 to prove Theorems 3 and 4.

Second Proof of Theorem 3: Note that the processes depicting the number of customers being served satisfy the hypotheses of Theorem 5. In the case of two separate facilities, the sum is not a birth-and-death process because the rates depend not only on the total number but how many are in the individual facilities. When k customers are present, the death (service) rates are identical, but the birth (arrival) rates can be higher in the combined system because if one of the separate facilities is full, then it cannot accept any more arrivals. Hence, the hypotheses of the corollary to Theorem 5 are satisfied. □

Proof of Theorem 4: Again we apply the corollary to Theorem 5. The reasoning is similar except here when k customers are present, the birth (arrival) rates are always identical, but the death (service) rates can be less with the separate facilities because there can be idle servers in one facility while there are customers waiting in the other facility. □

III. SAMPLE PATH COMPARISONS

Let $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ be real-valued stochastic processes. We call a real-valued function f defined on the space of all sample paths of $X(t)$ and $Y(t)$ nondecreasing if $f(\{x(t), t \geq 0\}) \leq f(\{y(t), t \geq 0\})$ for all sample paths $\{x(t), t \geq 0\}$ and $\{y(t), t \geq 0\}$ such that $x(t) \leq y(t)$ for all $t \geq 0$. We say the stochastic process $\{X(t), t \geq 0\}$ is stochastically less than or equal to the stochastic process $\{Y(t), t \geq 0\}$ and write $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$ if $f(\{X(t), t \geq 0\}) \leq_{st} f(\{Y(t), t \geq 0\})$ for all nondecreasing real-valued functions f defined on the sample paths of $X(t)$ and $Y(t)$. Clearly, stochastic order of the processes implies $X(t) \leq_{st} Y(t)$ for each t [just use the projection: $f(\{x(u), u \geq 0\}) = x(t)$], but it is much stronger, applying to many other nondecreasing functionals. In fact, since the queuing processes have sample paths with left and right limits everywhere, stochastic order of the processes is equivalent to stochastic order for all finite-dimensional (joint) distributions; see Section 4 of Kamae, Krengel, and O'Brien.[14] Moreover, stochastic order of the processes here is equivalent to the possibility of a strong sample-path comparison. In particular,

$$\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$$

holds if and only if it is possible to construct stochastic processes $\{\tilde{X}(t), t \geq 0\}$ and $\{\tilde{Y}(t), t \geq 0\}$ on a common probability space such that $\{\tilde{X}(t), t \geq 0\}$ has the same distribution as $\{X(t), t \geq 0\}$, $\{\tilde{Y}(t), t \geq 0\}$ has the same distribution as $\{Y(t), t \geq 0\}$, and every sample path of $\{\tilde{X}(t), t \geq 0\}$ lies below the corresponding sample path of $\{\tilde{Y}(t), t \geq 0\}$; see Theorem 1 of Kamae, Krengel, and O'Brien.[14] What we do is apply the easy half of this equivalence—the fact that the sample-path construction implies stochastic order—to make stochastic comparisons between the queuing processes. The proofs here are done by actually constructing processes with the sample-path ordering. Previous uses of such constructions appear in Sonderman,[7] Whitt,[8] Wolff,[9] and references therein. The approach is also closely related to the so-called "coupling" techniques; see Lindvall[15] and references therein.

We begin with the generalization of Theorem 2 for delay systems because it follows directly from Wolff.[9] As before, we assume the FCFS discipline, but now we allow the arrival streams in the two separate systems to be arbitrary. We assume the service times are independent of the arrival processes and mutually independent and identically distributed, but they need not be exponentially distributed. Since the arrival process is assumed to be independent of the service-time sequence, the evolution of the arrival process cannot depend on the state of the system. This excludes finite-source models, for which counterexamples to the efficiency of sharing are easy to construct; for example, see page 1377 of Beneš.[16] Let $Q_i(t)$ be the number of customers in the ith system and let $Q(t)$ be the number of customers in the combined system at time t.

Theorem 6: (Wolff) If $Q_1(0) = Q_2(0) = Q(0) = 0$, *then* $\{Q(t), t \geq 0\} \leq_{st} \{Q_1(t) + Q_2(t), t \geq 0\}$.

Remarks: (*i*) Wolff[9] was actually interested in comparing the FCFS discipline with the cyclic assignment discipline in a single delay-system. He showed that the queue length process with the FCFS discipline is stochastically less than the queue length process in the same system with any other discipline. This result applies here because the two separate facilities can be interpreted as a single system with a special queue discipline: Just label the arrivals in the special system according to the stream from which they came and then assign them according to the FCFS discipline to one of the servers in the corresponding subgroup of servers.

(*ii*) We can obtain corresponding results if the systems are not empty initially. For more general initial conditions, we can assume appropriate stochastic order for the residual service times at $t = 0$.

(*iii*) Wolff[9] also obtained similar comparison results for other processes, all of which hold here too: the departure epochs, the number of customers in queue, the total work (in service time) in the system, and the total work in queue. By the sample-path construction, the stochastic order jointly holds for all these processes. See Theorem 8 here.

(*iv*) As a consequence of Theorem 6, $Q(t) \leq_{st} Q_1(t) + Q_2(t)$ for each t. With the general assumptions here, steady-state distributions need not exist, but if $Q_i(t)$ and $Q(t)$ converge in law to Q_i and Q, respectively, as $t \rightarrow \infty$, then $Q \leq_{st} Q_1 + Q_2$; see Proposition 3 of Kamae, Krengel, and O'Brien.[14] The convergence of course holds in the setting of Theorem 2, so Theorem 6 implies (8) and thus Theorem 2.

(*v*) Since Theorem 2 concerns the mean-waiting time, it is natural to ask if the steady-state waiting-time distribution is also stochastically less in the combined system. Unfortunately, in general it is not. The counterexample in Whitt[17] applies here too; the cyclic discipline there can be interpreted as arrivals to separate facilities.

(*vi*) When the arrival streams are not Poisson, which we now permit, a new phenomenon occurs. Then the customers in the different streams experience different congestion when the systems are combined, even if the service times are independent and identically distributed. This phenomenon can be an important consideration in combining systems, but we do not consider it here; it has been studied by Kuczura.[18,19]

We now turn to our generalization of Theorem 1 for loss systems. In addition to allowing arbitrary arrival streams and general service-time distributions, we allow a finite waiting room. The number of waiting spaces in the combined system is the sum of the numbers of waiting spaces in the separate systems. Let $N_i(t)$ $[N(t)]$ be the number of customers lost in the interval $(0, t)$ in the ith separate system (in the combined system); let $S_i(t)$ $[S(t)]$ be the number of service completions in the interval $(0, t)$ in the ith separate system (in the combined system); and let $C_i(t)$ $[C(t)]$ be the amount of work performed—service given—in the interval $(0, t)$ in the ith separate system (in the combined system).

Theorem 7: If $Q_1(0) = Q_2(0) = Q(0) = 0$ *in these systems with finite waiting rooms, then*

$$\{N(t), t \geq 0\} \leq_{st} \{N_1(t) + N_2(t), t \geq 0\},$$

$$\{S(t), t \geq 0\} \geq_{st} \{S_1(t) + S_2(t), t \geq 0\},$$

and

$$\{C(t), t \geq 0\} \geq_{st} \{C_1(t) + C_2(t), t \geq 0\}.$$

Now assume that $N_i(t)/t$ and $N(t)/t$ converge (either in probability or with probability one) as $t \rightarrow \infty$. Let the limits be denoted $L(s_i, k_i, A_i(t), F)$ and $L(s_1 + s_2, k_1 + k_2, A_1(t) + A_2(t), F)$, respectively, with k_i denoting the number of waiting spaces, $A_i(t)$ the arbitrary arrival process and $F(x)$ the general service-time c.d.f. From Theorem 7 we immediately obtain the following generalization of Theorem 1.

Corollary: For all positive integers s_1, s_2, k_1 *and* k_2; *all arrival processes* $A_1(t)$ *and* $A_2(t)$; *and all service time c.d.f.'s* $F(x)$ *such that the loss-rate limits exist,*

$$L(s_1 + s_2, k_1 + k_2, A_1(t) + A_2(t), F) \leq L(s_1, k_1, A_1(t), F) + L(s_2, k_2, A_2(t), F).$$

To prove Theorem 7, we establish a finite-waiting-room generalization of Wolff's[9] comparison theorem. Following Wolff, we shall state the result in terms of the sample-path comparison. Since the joint distribution of the two systems being compared is artificially obtained, the appropriate conclusion is the general stochastic order as in Theorems 6 and 7.

We carry out the artificial construction by letting the systems being compared have identical arrival processes and service times. Note that we are now focusing on a single (arbitrary) sample path. We let the nth service time v_n be associated with the nth customer to enter service in each system rather than the nth arrival. Let a_n be the arrival epoch of the nth arrival, $0 \leq a_1 \leq a_2 \leq \cdots$. We assume there are s servers operating in parallel and k extra waiting spaces in both systems. We also assume the systems are initially empty.

One system, called the original system, will be the conventional system where the servers are fed by a single queue using a FCFS discipline. Moreover, there are k extra waiting spaces and arriving customers enter the system if the number of customers in the system is less than $s + k$, and are lost otherwise.

The other system, called the modified system, is any alternative to the original system which assigns customers to servers in some manner, independent of the sequence of service times $\{v_n\}$, and which loses arrivals whenever the system is full and in some manner otherwise.

Let a_n be the arrival epoch of the nth customer. For the original system, let t_n be the time that the nth customer to enter the system enters; obviously $t_n = a_k$ for some k, $k \geq n$. Also, for the original system, let b_n be the time the nth customer to begin service begins and let d_n be the nth ordered departure epoch from the system. Let a_n, t'_n, b'_n, and d'_n be the corresponding quantities for the modified system.

Theorem 8: For all integers n, $t_n \leq t'_n$, $b_n \leq b'_n$, and $d_n \leq d'_n$.

Proof: The sets of unordered departure epochs in the two systems are clearly $\{(b_n + v_n)\}$ and $\{(b'_n + v_n)\}$, respectively. For the original system,

$$d_1 = \min_{1 \leq i \leq s} \{b_i + v_i\} = \min_{i \leq i \leq s} \{t_i + v_i\},$$

$$d_n = n\text{th-order statistic from } \{(b_i + v_i): i < n + s\} \qquad (10)$$

and

$$b_n = \max\{t_n, d_{n-s}\}, \qquad n \geq 1, \qquad (11)$$

where $d_j = 0$ if $j = 0$. For the modified system,

$$d'_n = n\text{th-order statistic from } \{(b'_i + v_i): i < n + s\} \qquad (12)$$

and

$$b'_n \geq \max\{t_n, d'_{n-s}\}, \qquad n \geq 1, \qquad (13)$$

because in the modified system it is possible to have a positive queue and an idle server.

Since the nth-order statistic is a monotonic function, to prove Theorem 8 it suffices to show that $t_n \leq t'_n$ and $b_n \leq b'_n$ for all n. We show this induction. Obviously $b_i = t_i = a_i \leq t'_i \leq b'_i$, $1 \leq i \leq s$. Suppose $t_i \leq t'_i$ and $b_i \leq b'_i$ for all i, $i \leq n - 1$. We first show that $t_n \leq t'_n$. Suppose not; then

$$t_n > t'_n \geq t'_{n-1} \geq t_{n-1},$$

and thus $n - 1$ customers have entered both systems before the arrival associated with t'_n. (Note that customers could arrive in batches, i.e., $a_k = a_{k+1}$ is a possibility, but this presents no serious difficulty.) However, by the induction hypothesis $b_i + v_i \leq b'_i + v_i$, $i \leq n - 1$, so the original system has at most the same number of customers as the modified system before the arrival associated with t'_n. Thus, $t_n > t'_n$ cannot occur. Hence $t_n \leq t'_n$ as claimed.

To continue the induction proof for b_n, note that (10) and (12) imply that $d_i \leq d'_i$ for $i \leq n - s$. Then, from (11) and (13), we have

$$b_n = \max\{t_n, d_{n-s}\} \leq \max\{t'_n, d'_{n-s}\} \leq b'_n,$$

which completes the proof. □

Remark: Our proof of Theorem 8 is closely related not only to Wolff's proof,[9] but also to Sonderman's comparison proofs.[20,21] Sonderman was concerned with the effect of different service-time distributions instead of different queue disciplines.

We close this section with another result about pure-loss systems. With waiting rooms or with general service times it is easy to show that the stochastic processes representing the number of customers in the system need not be stochastically ordered, but we do get stochastic order with exponential distributions and no waiting rooms.

Theorem 9: In the setting of Theorem 7, if there are no waiting rooms, if the service-time distribution is exponential and if $Q(0) =_{st} Q_1(0) + Q_2(0)$, *then*

$$\{Q(t),\, t \geq 0\} \geq_{st} \{Q_1(t) + Q_2(t),\, t \geq 0\}$$

and

$$\{N(t),\, t \geq 0\} \leq_{st} \{N_1(t) + N_2(t),\, t \geq 0\}.$$

Proof: Here the argument follows Sonderman[7,20,21] and Whitt.[8] As the first step in constructing the two systems on the same probability space, we let the two systems being compared have identical arrival processes; i.e., we let the arrival process to the combined group of $s_1 + s_2$ servers be the sum of the two arrival processes to the separate groups of s_i servers. This not only means that the arrival processes have the same joint distributions, but that they have the same sample paths. Similarly, we let both systems start off with the same number of customers in the system; i.e., given the pair $[Q_1(0), Q_2(0)]$, we let $Q(0) = Q_1(0) + Q_2(0)$. We now show how to construct the departures so that

$$N(t) \leq N_1(t) + N_2(t) \tag{14}$$

and

$$Q(t) \geq Q_1(t) + Q_2(t) \tag{15}$$

for all $t \geq 0$. We generate departures from both systems using a single Poisson process with rate $(s_1 + s_2)\mu$. Each point in this Poisson process corresponds to a potential departure. Suppose the point occurs at time t. With probability $Q_1(t)/(s_1 + s_2)$, the point corresponds to a departure from both the single group of s_1 servers and the combined group of $s_1 + s_2$ servers; with probability $Q_2(t)/(s_1 + s_2)$, the point corresponds to a departure from both the single group of s_2 servers and the combined group of $s_1 + s_2$ servers; with probability $[Q(t) - Q_1(t) - Q_2(t)]/(s_1 + s_2)$ the point corresponds to a departure from only the combined group of $s_1 + s_2$ servers; and finally, with probability $[s_1 + s_2 - Q(t)]/(s_1 + s_2)$, the point corresponds to no departure at all. This can be shown to yield the proper distributions for each system; see Sonderman[7] for more detail. This also guarantees that there is a departure in the combined group of $s_1 + s_2$ servers whenever there is a departure from one of the groups with s_1 and s_2 servers. There also cannot be a departure from the combined group alone when $Q(t-) = Q_1(t-) + Q_2(t-)$, so inequality (15) is maintained. This means that all departures and losses from the combined group of $s_1 + s_2$ servers that are not matched by corresponding departures or losses from one of the groups of s_1 and s_2 servers can be matched with earlier losses from one of the groups of s_1 and s_2 servers. Mathematical induction on the arrival index establishes (14) and (15) and formally completes the proof. □

Remark: For the special case of $M/M/s$ systems, the stochastic order in Theorems 6 and 9 can also be established under the conditions in the corollary to Theorem 5 using existing comparison theorems for continuous-time Markov chains; see Sonderman.[7] However, we know of no direct connections between the monotone likelihood-ratio orderings and the sample-path orderings.

IV. DIFFERENT SERVICE RATES

In this section we let the service rates in the two separate systems be different. One way to extend (2) and (4) occurs when the service times are associated with the arrivals. If two independent Poisson streams with rates λ_1 and λ_2 and associated service-time c.d.f.'s $F_1(x)$ and $F_2(x)$ are combined, then the resultant stream is a Poisson stream with rate $\lambda_1 + \lambda_2$ and associated service-time c.d.f.:

$$F(x) = [\lambda_1 F_1(x) + \lambda_2 F_2(x)]/(\lambda_1 + \lambda_2).$$

Of course, when $F_i(x)$ is exponential with mean μ_i^{-1} for each i, $F(x)$ is not exponential unless $\mu_1 = \mu_2$. However, the blocking probability for an $M/G/s$ loss system depends only on the mean service time. Thus the loss rate for the combined system is $L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2))$, and a natural extension of (2) to conjecture is

$$L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \le L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2).$$

Unfortunately, this conjectured inequality is not valid. To see this, let $\lambda_1 = 1$, $\mu_1 = \epsilon^{-1}$, $\lambda_2 = \epsilon$, and $\mu_2 = \epsilon^2$; then $a_1 = \epsilon$ and $a_2 = \epsilon^{-1}$. Obviously,

$$\begin{aligned} L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) &= (\lambda_1 + \lambda_2)B(s_1 + s_2, a_1 + a_2) \\ &= (1 + \epsilon)B(s_1 + s_2, \epsilon + \epsilon^{-1}) \\ &\rightarrow 1 \quad \text{as} \quad \epsilon \rightarrow 0, \end{aligned}$$

whereas

$$\begin{aligned} L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2) &= \lambda_1 B(s_1, a_1) + \lambda_2 B(s_2, a_2) \\ &= B(s_1, \epsilon) + \epsilon B(s_2, \epsilon^{-1}) \\ &\rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0. \end{aligned}$$

Consequently, in this case

$$L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \geq L(s_1, \lambda_1, \mu_1) + L(s_2, \lambda_2, \mu_2)$$

for sufficiently small ϵ. The previous measure, rate of customer loss, is not the only reasonable way to evaluate system performance in this case. For example, one might be interested in the rate of loss of service time. (Note that there is no real difference between these measures when the mean service times of the systems are identical.) With this new measure, the natural extension of (2) to conjecture is:

$$\frac{a_1 + a_2}{\lambda_1 + \lambda_2} L(s_1 + s_2, \lambda_1 + \lambda_2, (\lambda_1 + \lambda_2)/(a_1 + a_2)) \leq \frac{1}{\mu_1} L(s_1, \lambda_1, \mu_1) + \frac{1}{\mu_2} L(s_2, \lambda_2, \mu_2).$$

This inequality is in fact always true, since substitution of $L(s, \lambda, \mu) = \lambda B(s, \lambda/\mu)$ quickly reduces it to the second version of the inequality of Theorem 1. Thus the server occupancy is always increased for the combined system.

Turning to delay systems, we again find examples where sharing can be counterproductive. To see this, consider two $M/M/1$ delay systems with $\lambda_1 = 1$, $\mu_1 = 2$, $\lambda_2 = \epsilon$, and $\mu_2 = 2\epsilon$. Then $EQ_1(\infty) = EQ_2(\infty) = \rho/(1 - \rho) = 1$, but $EQ(\infty)$ can be shown to be of order ϵ^{-1} as $\epsilon \rightarrow 0$: Consider the interval following a low-intensity arrival. With probability $\lambda_2/(\lambda_2 + \mu_2) = 1/3$, a second low-intensity arrival occurs before the first departs. Then there follows an exponentially distributed interval of mean length $1/4\epsilon$ during which the combined system fills up with high-intensity customers. In computing the average number of customers in the system, we get a term of order ϵ^{-2} (the total area in the plot of the number of customers in the system versus time, starting from the moment the second low-intensity customer arrives and ending when one of the two low-intensity customers departs), divided by a term of order ϵ^{-1}. In other words, with the mean steady-state delays held fixed in the two separate systems, the mean steady-state delay in the combined system can be arbitrarily large.

Note that the combined system can be modeled as an $M/G/s_1{+}s_2$ delay system where the service-time distribution is the mixture of two exponential distributions, but in contrast to the case of loss systems the mean delay does not depend only on the mean of the service-time distribution. Hence, the appropriate generalization of (4) involves a system which is not $M/M/s$.

Another possible extension for $\mu_1 \neq \mu_2$ occurs when the service-time distributions are associated with the servers. Here the combined system is not $M/M/s$ because there are heterogeneous servers, so there are no equations similar to (2) and (4). In this case, it can be shown that with exponential service-time distributions and no waiting rooms, resource sharing is always better if customers always are sent to the fastest available server. In particular, as in Theorems 6 to 8, it can be shown for any single system that assigning customers to the fastest available server produces fewer losses than any other rule, where by "fewer losses" we mean in the sample-path ordering of Section III. One other rule, corresponding to the two separate systems, is to assign the customer only to servers associated with their original separate arrival streams.

When we focus on delay systems with heterogeneous servers, it is easy to give counterexamples showing that resource sharing can again be counterproductive. Related literature on the assignment of customers to heterogeneous servers appears in Winston,[22-24] Smith,[25] and references therein.

This section shows that, with unequal service-time distributions, resource sharing can be counterproductive. However, with unequal service-time distributions, much depends on the criterion of system performance. Also, it should be noted that such counterexamples have been observed before; others have discovered that infrequent "bad" customers can affect adversely a large number of "good" customers.

V. ACKNOWLEDGMENT

We are grateful to Ed Arthurs and Bart Stuck for suggesting this problem.

APPENDIX

Here we give two results that are due to others. First, we present Paul Burke's proof that $B(ts, ta)$ is strictly decreasing in t for $t \geq 0$. This result implies Theorem 1 when $\lambda_1/s_1 = \lambda_2/s_2$, because then $B(s_1 + s_2, a_1 + a_2) = B(ts_i, ta_i)$ for some $t \geq 1$, so $B(s_i, a_i) \geq B(s_1 + s_2, a_1 + a_2)$ for each i and

$$\frac{\lambda_1 B(s_1, a_1)}{\lambda_1 + \lambda_2} + \frac{\lambda_2 B(s_2, a_2)}{\lambda_1 + \lambda_2} \geq B(s_1 + s_2, a_1 + a_2),$$

which is equivalent to (2).

To see that $B(ts, ta)$ is strictly decreasing in t, first recall the following equation relating two different expressions for the tail of the gamma distribution:

$$e^{-a}\sum_{k=0}^{k=n} a^k/k! = \int_a^\infty \frac{x^n e^{-x}}{n!}\,dx.$$

Then note that

$$\frac{1}{B(ts,\,ta)} = \frac{\displaystyle\int_{ta}^\infty e^{-x}x^{ts}\,dx}{e^{-ta}(ta)^{ts}}$$

$$= \int_{ta}^\infty e^{-(x-ta)}(x/ta)^{ts}\,dx$$

$$= \int_0^\infty e^{-x}(1+[x/ta])^{ts}\,dx;$$

also see Theorem 3 of Jagerman.[3] Finally, $(1+[x/ta])^{ts}$ is strictly increasing in t.

Second, Herbert Shulman has shown that Paul Burke's result and the convexity of $B(s, a)$ in s for $s \geq 1$ imply a version of Theorem 1. Such convexity has frequently been conjectured but has been proved only for lattices of points with unit spacing, see Messerli[26] and references therein. These versions of convexity are not strong enough to make the following valid even when s_1 and s_2 are integers; however, general convexity would establish the proof for all real numbers s_1 and $s_2 \geq 1$, a more general mathematical result. We reproduce Shulman's argument here:

$$B(s_1+s_2,\,a_1+a_2) \leq \frac{a_1}{a_1+a_2}B\left(\frac{a_1+a_2}{a_1}s_1,\,a_1+a_2\right)$$

$$+\frac{a_2}{a_1+a_2}B\left(\frac{a_1+a_2}{a_2}s_2,\,a_1+a_2\right)$$

$$\leq \frac{a_1}{a_1+a_2}B(s_1,\,a_1) + \frac{a_2}{a_1+a_2}B(s_2,\,a_2).$$

REFERENCES

1. E. Arthurs and B. W. Stuck, "Subadditivity of Teletraffic Special Functions," SIAM Rev., to be published.
2. L. Kleinrock, *Queueing Systems, Volume 1: Theory*, New York: John Wiley, 1975.
3. D. L. Jagerman, "Some Properties of the Erlang Loss Function," B.S.T.J., *53*, No. 3 (March 1974), pp. 525–51.
4. R. B. Cooper, *Introduction to Queueing Theory*, New York: Macmillan, 1972.
5. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, New York: John Wiley, 1976.
6. S. Stidham, Jr., "A Last Word on $L = \lambda W$," Oper. Res., *22*, No. 2 (March–April 1974), pp. 417–22.

7. D. Sonderman, "Comparing Semi-Markov Processes," Math. Oper. Res. *5*, No. 1 (February 1980), pp. 110–20.
8. W. Whitt, "Comparing Counting Processes and Queues," Adv. Appl. Probab., *13*, No. 1 (March 1981), to be published.
9. R. W. Wolff, "An Upper Bound for Multi-Channel Queues," J. Appl. Probab., *14*, No. 4 (December 1977), pp. 884–8.
10. T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*, New York: Academic Press, 1967.
11. W. Whitt, "A Note on the Influence of the Sample on the Posterior Distribution," J. Am. Statist. Assoc., *74*, No. 366 (June 1979), pp. 424–6.
12. W. Whitt, "Uniform Conditional Stochastic Order," J. Appl. Probab., *17*, No. 1 (March 1980), pp. 112–23.
13. J. Keilson and U. Sumita, "Uniform Stochastic Ordering," Working Paper, The Graduate School of Management, University of Rochester, 1980.
14. T. Kamae, U. Krengel, and G. L. O'Brien, "Stochastic Inequalities on Partially Ordered Spaces," Ann. Probab., *5*, No. 6 (December 1977), pp. 899–912.
15. T. Lindvall, "A Note on Coupling of Birth and Death Processes," J. Appl. Probab., *16*, No. 3 (September 1979), pp. 505–12.
16. V. E. Beneš, "Programming and Control Problems Arising from Optimal Routing in Telephone Networks," B.S.T.J., *45*, No. 9 (November 1966), p. 1373.
17. W. Whitt, "On Stochastic Bounds for the Delay Distribution in the *GI/G/s* Queue," Oper. Res., to be published.
18. A. Kuczura, "Queues with Mixed Renewal and Poisson Inputs," B.S.T.J., *51*, No. 6 (July–August 1972), pp. 1305–26.
19. A. Kuczura, "Loss Systems with Mixed Renewal and Poisson Inputs," Oper. Res., *21*, No. 3 (May–June 1973), pp. 787–95.
20. D. Sonderman, "Comparing Multi-Server Queues with Finite Waiting Rooms, I: Same Number of Servers," Adv. Appl. Probab., *11*, No. 2 (June 1979), pp. 439–47.
21. D. Sonderman, "Comparing Multi-Server Queues with Finite Waiting Rooms, II: Different Numbers of Servers," Adv. Appl. Probab., *11*, No. 2 (June 1979), pp. 448–55.
22. W. L. Winston, "Assignment of Customers to Servers in a Heterogeneous Queueing System with Switching," Oper. Res., *25*, No. 3 (May–June 1977), pp. 468–83.
23. W. L. Winston, "Optimal Dynamic Rules for Assigning Customers to Servers in a Heterogeneous Queueing System," Nav. Res. Logis. Q., *24*, No. 2 (June 1977), pp. 293–300.
24. W. L. Winston, "Optimal Assignment of Customers in a Two-Server Congestion System with No Waiting Room," Manage. Sci., *24*, No. 6 (February 1978), pp. 702–5.
25. D. R. Smith, "Optimal Repair of a Series System," Oper. Res., *26*, No. 4 (July–August 1978), pp. 653–62.
26. E. J. Messerli, "Proof of a Convexity Property of the Erlang B Formula," B.S.T.J., *51*, No. 4 (April 1972), pp. 951–3.