AT&T Technical Journal
Vol. 64, No. 8, October 1985
Printed in U.S.A.

Blocking When Service Is Required From Several Facilities Simultaneously

By W. WHITT*

(Manuscript received January 14, 1985)

This paper analyzes a mathematical model of a blocking system with simultaneous resource possession. There are several multiserver service facilities without extra waiting space at which several classes of customers arrive in independent Poisson processes. Each customer requests service from one server in each facility in a subset of the service facilities, with the subset depending on the customer class. If service can be provided immediately upon arrival at all required facilities, then service begins and all servers assigned to the customer start and finish together. Otherwise, the attempt is blocked (lost without generating retrials). The problem is to determine the blocking probability for each customer class. An exact expression is available, but it is complicated. Hence, this paper investigates approximation schemes.

I. INTRODUCTION AND SUMMARY

The multifacility blocking problem considered here arises in many contexts and has a long history in traffic engineering (see pages 77 and 95 of Ref. 1). We were motivated by performance analysis issues in packet-switched communication networks. Specifically, we were investigating methods for calculating the blocking probabilities (percentage of failed attempts) in setting up virtual circuits. The need for methods to calculate these blocking probabilities arose in the development of the Packet Network Performance Analysis module of the Packet Network Design and Analysis (PANDA) software package in

* AT&T Bell Laboratories.

the Operations Research Department of AT&T Bell Laboratories.[2,3] It is difficult to analyze the blocking because a circuit typically requires the simultaneous possession of limited resources associated with several different facilities (transmission links, memory buffers, etc.). Moreover, there is competition for the resources not only from other demands for circuits on the same path, but also from demands for different circuits that use only some of the same facilities. Hence, even without alternate routing or waiting (which we do not consider), the blocking is complicated. Our purpose here is to develop bounds and approximations. After we describe our model, we will discuss related work and other applications.

1.1 The mathematical model

There are n multiserver service facilities without extra waiting room and c customer classes. Service facility i has s_i servers. Customers from class j arrive according to a Poisson process with rate λ_j and immediately request service from one server at each facility in a subset A_j of the n service facilities. If all servers are busy in any of the required facilities, the request is blocked (lost without generating retrials). Otherwise, service begins immediately in all the required facilities. All servers working on a given customer from class j start and free up together. The service time for class j at all facilities has a general distribution with finite mean μ_j^{-1}. We assume that the c arrival processes and all the service times are mutually independent.

This model already embodies the extension in which each class requires service from a random subset of the n facilities. Suppose that class j with arrival rate λ_j initially requires one server at each facility in subset A_{jk} with probability p_{jk}, where $\sum_k p_{jk} = 1$. We can represent this more general model within our framework by increasing the number of classes. New class (j, k) has a Poisson arrival process with rate $\lambda_j p_{jk}$ and requires one server in each facility in the subset A_{jk}. This procedure is justified because of two familiar properties of Poisson processes: (1) independent random splitting of a Poisson process produces independent Poisson processes, and (2) the superposition of independent Poisson processes is a Poisson process (see Theorems 4.2 and 5.3 of Ref. 4).

Returning to the previous setting in which each class requires a fixed subset of facilities, we let $b(A)$ be the probability that all servers are busy in at least one facility in the subset A (at an arbitrary time in steady state). Thus $b(i) \equiv b(\{i\})$ is the probability all servers are busy at facility i. Since Poisson arrivals see time averages,[5] $b(A_j)$ is also the blocking probability for class j. [The blocking probability for class j would be $\sum_k p_{jk} b(A_{jk})$ if class j required a random number of facilities, as described above.]

It is not difficult to give the exact formula for $b(A)$ using theory related to queueing networks,[6-8] but the formula is complicated, especially when the numbers n and c are large (see Theorem 4 in Section II). To appreciate the complexity, recall that the arrival rates λ_j, service rates μ_j, and subsets of required facilities A_j for class j can differ from class to class. Hence, our interest centers on developing bounds and approximations.

1.2 Related work

There is a substantial body of related literature. The problem treated here is connected, at least in spirit, to the theory of gradings and link systems in traffic engineering.[1] The specific approximation problem is discussed by Holtzman.[9] Also somewhat related is the work on stochastic models of dynamic storage allocation.[10-12] Previous work also has been done on service systems, with waiting as well as blocking, in which customers require more than one server.[7-20]

Our model is relatively elementary compared with many of these other models. Our analysis benefits by having blocking instead of waiting and by having each customer require exactly one server per facility. On the other hand, we address an issue typically not considered in the papers in which customers require more than one server: Here there are constraints on which servers can be used; there must be one server from each facility. To make the comparison clear, it is useful to modify our model by considering one large facility containing all the servers in all the original facilities. If one of our customers requiring service from one server in each of k facilities could use any k servers in the single large facility, then we would have the model of Arthurs and Stuck.[16] Here, however, there are constraints.

The model considered here is in fact a special case of a more general single-facility blocking model of Kaufman,[7] in which there is a general sharing rule. From Kaufman or Burman, Lehoczky and Lim,[8] we learn that our model is a product-form model with the insensitivity property.[6] This provides expressions for the exact blocking probabilities (Theorem 4 in Section 2.1 here), but as noted above this exact expression is quite intractable. The insensitivity property tells us that the blocking probabilities depend on the service-time distributions only through their means, so that there is no need to assume exponential service-time distributions; for convenience we can replace general service-time distributions by exponential service-time distributions without affecting the blocking probabilities. We discuss insensitivity further in Sections 1.8 and IV.

A special case of our model has also been analyzed in a database locking study by Mitra and Weinberger.[21,22] In their model, the facilities are items in the database and the customers are transactions that

"touch" a specific set of items. To maintain consistency, only one transaction is allowed to touch each item at any time, so that in their model transactions requiring items already being touched are blocked. Their model thus corresponds to the special case of our multifacility blocking model in which each facility has a single server. (It may be of interest to consider the extension of their model to multiserver facilities to represent multiple copies of database items in the database.)

Mitra and Weinberger show that the analysis can be greatly simplified by focusing attention on special symmetric versions of the model. They assume that the arrival rate and service rate for each customer class that requires k facilities (items) is independent of the subset of k facilities required. Moreover, they assume that there is a customer class for each subset of size k. Most important, they consider only the case of one server per facility. (It should be clear that the case of multiserver facilities is much harder.) For these special symmetric models, they obtain an efficient algorithm for calculating the partition function of the product-form model, from which the desired blocking probabilities are easily obtained. [For some database locking applications, it may not be reasonable to assume that the arrival rates for all subsets of size k are identical. Then the approximation methods in this paper may be helpful. See Remark 3 in Section 1.6.]

The symmetric case of the multifacility blocking model has also been considered by Heyman in the investigation of a communication system.[23] We shall also discuss symmetric models here, beginning in Section 1.6. For symmetric models, the approximations are more reliable and much easier to compute.

We have mentioned that this work was primarily motivated by performance analysis issues in packet-switched communication networks, specifically in the PANDA software package.[2,3] The approximations here have also been applied to study the blocking in an AT&T Bell Laboratories computer network[24] and an AT&T Communications model for overseas voice traffic.[25] Another example of the multifacility blocking problem in telephony is contained in Akinpelu.[26] The work that bears most directly on this paper is in Refs. 2, 3, 7, 8, 9, 21, and 23 through 26. (Also see Section VIII.)

1.3 Summary and organization of this paper

We describe our main results in the rest of Section I and provide the supporting technical details in the remaining sections. We discuss three different approximation schemes: the summation bound, the product bound, and the reduced-load approximation. The two bounds are well-known approximations. The reduced-load approximation evidently has a long history,[9] but is not as well known as it deserves to

be. We propose for the reduced-load approximation a successive approximation scheme that is very easy to implement and seems to perform well. In particular, the reduced-load approximation with the successive approximation scheme is ideally suited for large models, where the exact formula becomes intractable.

We obtain two major results about these approximation schemes: (1) As suggested by the names, the first two approximation schemes indeed yield upper bounds on the blocking probabilities, and (2) a limit theorem establishes that the third approximation scheme, the reduced-load approximation, is asymptotically correct for symmetric models as the size of the model grows, in a sense which we will make precise. It is significant that the limiting conditions do not correspond to light traffic as in Refs. 21 and 23, so that in this limit the reduced-load approximation can be very different from the bounds. Our two main results have mathematical interest as well as applied interest, because they are obtained by focusing on multidimensional stochastic processes that are not Markov.

We also obtain two additional light-traffic results. First, we show that all three approximations are asymptotically equivalent as the loads decrease (Corollary 2.3). Second, we show that all three approximations are asymptotically correct as the loads decrease for symmetric models (Corollary 3.2). The qualification "for symmetric models" in the last sentence is important because it can happen for asymmetric models that all three approximations are equally bad in light traffic (see the remark at the end of Section 1.5). As we mentioned in Section 1.2, the approximations are more reliable and easier to use with symmetric models, but we believe they are also very useful for asymmetric models when applied with some caution.

We discuss the bounds in Section 1.4, the reduced-load approximation in Section 1.5, and the main limit theorem in Section 1.6. We discuss numerical examples in Section 1.7; existence, uniqueness, and insensitivity of equilibrium blocking probabilities in Section 1.8; and an extension of the reduced-load approximation for non-Poisson arrival processes in Section 1.9. Additional technical details will be provided in subsequent sections. The main results and directions for future research are summarized in Section VI.

Here are the principal conclusions from our analysis and limited numerical experience: For light loads, for example, blocking in the order of 0.01 or less, the elementary bounds are usually adequate approximations for engineering purposes. Having established that these approximations are bounds, we gain some peace of mind in knowing that the approximations are conservative. For higher levels of blocking, such as 0.05 and above, the reduced-load approximation typically does much better than the elementary bounds. Moreover, the

successive approximation scheme proposed for the reduced-load approximation is very easy to use (Theorem 2), so that the reduced-load approximation seems very attractive when the loads need not be light.

1.4 Bounds

Let $B(s, \alpha)$ be the classical Erlang blocking formula associated with the M/G/s/loss service system with s servers and offered load α,[27,28] defined by

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

where, as usual, the offered load α is the arrival rate multiplied by the expected service time. Let C_i be the set of all classes that request service from facility i, that is,

$$C_i = \{j: i \in A_j\}. \tag{2}$$

Let $\hat{\alpha}_i$ be the offered load at facility i (not counting blocking elsewhere), defined by

$$\hat{\alpha}_i = \sum_{j \in Ci} \alpha_j, \tag{3}$$

where $\alpha_j = \lambda_j/\mu_j$ is the offered load of class j to the system as a whole.

In Section II we establish the following bounds. These bounds are standard approximations that have long been regarded as conservative.[1] We show that intuition is correct in this case.

Theorem 1: (Product Bound) For each subset A,

$$b(A) \leq 1 - \prod_{i \in A} [1 - B(s_i, \hat{\alpha}_i)].$$

Corollary 1.1: (Facility Bound) For each i, $b(i) \leq B(s_i, \hat{\alpha}_i)$.

Proof: Let $A = \{i\}$. □

Corollary 1.2: (Summation Bound) For each subset A,

$$b(A) \leq \sum_{i \in A} B(s_i, \hat{\alpha}_i).$$

Proof: The summation bound in Corollary 1.2 is always greater than or equal to the product bound in Theorem 1, as is easily verified by induction on the number of facilities in A. Corollary 1.2 also follows directly from Corollary 1.1 and the Bonferroni inequalities (see page 110 of Ref. 29). □

Remarks: For the special case of two facilities, Corollary 1.2 has been proved by different methods by D. D. Sheng and D. R. Smith; see the appendix in Ref. 3. The simple approximation provided by the summation bound in Corollary 1.2 was used in early versions of the PANDA software package,[2] before being replaced by the product bound

in Theorem 1.[3] The summation bound in Corollary 1.2 coincides with the asymptotic approximation developed by Mitra and Weinberger,[21] which is a light-traffic limit. (See Corollaries 2.3 and 3.2 below.)

We give a separate proof of the facility bound in Corollary 1.1, which is of independent interest. We apply Theorem 5 of Smith and Whitt[30] to obtain a monotone-likelihood-ratio ordering for the number of busy servers (Theorem 5), which does not follow from our proof of Theorem 1.

Our proof of Theorem 1 is based on a general technique for comparing a non-Markov process to a Markov process using the conditional transition rates, which applies to many different definitions of stochastic order (Theorem 6). We apply this technique to prove Theorem 1 using the version of stochastic order for probability distributions on R^n based on comparing cumulative distribution functions (Theorem 7). For $n > 1$, this stochastic ordering is weaker than the standard form of stochastic order based on all increasing sets. Our general approach for comparing a non-Markov process to a Markov process has much wider applicability, and is discussed further elsewhere.[31] Our approach exploits stochastic monotonicity of the Markov process,[32-35] and is closely related to the stochastic comparisons for multidimensional Markov processes by Massey that have been applied to establish comparisons for Markovian queueing network models.[36-38]

The bound in Theorem 1 corresponds to independent blocking in the different facilities with the bound Corollary 1.1 used in each facility. It is natural to conjecture that Theorem 1 might be obtained from the more easily established Corollary 1.1 and the inequality

$$b(A) \leq 1 - \prod_{j \in A} [1 - b(i)] \tag{4}$$

but (4) is *not* valid in general (see Example 6 in Section II).

For typical applications in which the bounds are relatively small and customers require only a few facilities, the bounds usually are excellent approximations (see Corollary 3.2), but the following examples demonstrate that the bounds are not always good approximations.

Example 1: Suppose that all n facilities have s servers. Let there be only one customer class, which requests service from all n facilities. Then $\hat{\alpha}_1 = \alpha_1$ and $b(\{1, \cdots, n\}) = b(1) = B(s, \alpha_1) = B(s, \hat{\alpha}_1)$, so that the bound in Corollary 1.1 is tight (an equality), but the bounds in Theorem 1 and Corollary 1.2 can be poor approximations. □

Example 2: Suppose that there are two facilities and two customer classes. Let $s_1 = 10$, $s_2 = 1$, $A_1 = \{1\}$, $A_2 = \{1, 2\}$, $\alpha_1 = 1$, and $\alpha_2 = 100$. Then $B(s_1, \hat{\alpha}_1) = B(10, 101) \approx 1$, but $b(1) \approx 0$, because at most one class 2 customer can be in service at any time. Hence, in this case the bound in Corollary 1.1 is a very bad approximation. □

1.5 A reduced-load approximation

Since the approximations in Theorem 1 and its corollaries are upper bounds, it is natural to look for reduced values that might be better approximations. One way to do this is to reduce the offered load $\hat{\alpha}_i$ at facility i by taking into account the blocking elsewhere. It is natural to develop such an approximation within a framework of facility independence, that is, the assumption that the events of blocking at the difficult facilities are independent. This seems to be a reasonable approximating assumption for "typical" examples, which has been applied before for multiple facilities.[1,14] As a consequence, we have the facility-independence approximation

$$1 - b(A) \approx \prod_{i \in A} [1 - b(i)]. \tag{5}$$

Next we introduce the following approximate total offered load at facility i using the facility independence assumed above:

$$\bar{\alpha}_i = \sum_{j \in C_i} \alpha_j \prod_{\substack{k \in A_j \\ k \neq i}} [1 - b(k)]. \tag{6}$$

In (6) we have reduced the offered load α_j of class j at facility i by the blocking elsewhere. Of course, using (6) we make the offered loads dependent on the blocking probabilities as well as vice versa. [However, the facility-independence approximation in (5) greatly reduces the complexity.] Hence, this leads to a system of equations characterizing the blocking probabilities as our approximation. In particular, our proposed reduced-load approximation for the blocking probability at facility i is the solution to the following system of equations:

$$b^*(i) = B\left\{s_i, \sum_{j \in C_i} \alpha_j \prod_{\substack{k \in A_j \\ k \neq i}} [1 - b^*(k)]\right\}, \qquad 1 \leq i \leq n. \tag{7}$$

From (1), we see that (7) yields n polynomial equations in the n unknowns $b^*(1), \cdots, b^*(n)$.

In general, solving a system of n nonlinear equations in n unknowns can be quite unpleasant. Of course, in many situations there are symmetries in the model, which allow us to reduce the number of equations (and variables). In fact, in the next section we discuss the totally symmetric model, for which (7) reduces to one equation in one variable, which is trivial to solve. (The database model in Ref. 21 also simplifies in this way.) However, we also propose a relatively simple computational scheme for solving the general system (7). In particular, we suggest using successive approximations, that is, iteratively applying the right side of (7) to successive candidate vectors of blocking probabilities. The following theorem indicates that (7) always has a

solution and that the successive approximation scheme either finds the unique solution or provides upper and lower bounds on all solutions to (7).

Theorem 2: (Existence and Successive Approximation) If s_i *and* $\hat{\alpha}_i$ *are strictly positive for each* i*, then the system of eqs.* (7) *has a solution* $\mathbf{b}^* \equiv [b^*(1), \cdots, b^*(n)]$ *with* $0 < b^*(i) < 1$ *for all* i*. All solutions* $\mathbf{b}^*$ *can be bounded above and below, and sometimes found, by successive approximation, that is, by iteratively applying the operator* $T \equiv T\{[b(1), \cdots, b(n)]\}$ *mapping* $[0, 1]^n$ *into itself defined by the right side of* (7)*, starting with* $\mathbf{1} \equiv (1, 1, \cdots, 1)$*. In particular, successive applications of* T *yield the following upper and lower bounds on* $[b^*(1), \cdots, b^*(n)]$ *for all* k*:*

$$(0, 0, \cdots, 0) \equiv \mathbf{0} = T(\mathbf{1}) < T^{2k+1}(\mathbf{1}) < T^{2k+3}(\mathbf{1}) < [b^*(1), \cdots, b^*(n)] < T^{2k+2}(\mathbf{1}) < T^{2k}(\mathbf{1}) < T^0(\mathbf{1}) = \mathbf{1} \equiv (1, 1, \cdots, 1). \quad (8)$$

Proof: First, the operator T defined by the right side of (7) obviously maps $[0, 1]^n$ into itself. Since T is continuous, T has a fixed point, by the classical Brouwer fixed point theorem.[39] Let $\mathbf{b}^*$ represent such a fixed point. Since the operator T is strictly decreasing, $b(i) > b^*(i) > T(\mathbf{b})_i$ for all i, where $\mathbf{b} \equiv (b(1), \cdots, b(n))$, whenever $b(i) > T(\mathbf{b})_i$ for all i and $b(i) < b^*(i) < T(\mathbf{b})_i$ for all i whenever $b(i) < T(\mathbf{b})_i$ for all i. Since $T(\mathbf{1}) = \mathbf{0}$ and $T(\mathbf{0}) < \mathbf{1}$, $0 < b^*(i) < 1$ for all i. □

Since T is continuous and strictly decreasing, the iterated operator $T^{(2)}$ is continuous and strictly increasing. Hence, the successive approximation scheme (8) converges in the sense that $T^{2k+1}(\mathbf{1}) \to \mathbf{L}$ and $T^{2k}(\mathbf{1}) \to \mathbf{U}$, where $\mathbf{L} = (L_1, \cdots, L_n)$ and $\mathbf{U} = (U_1, \cdots, U_n)$ are lower and upper bonds, respectively, on any solution to (7), that is, $L(i) \leq b^*(i) \leq U(i)$, $1 \leq i \leq n$. Often we will have $\mathbf{L} = \mathbf{U}$, that is, $L(i) = U(i) = b^*(i)$ for all $\mathbf{i}$, but not always, because $T^{(2)}$ can have more than one fixed point, as Example 3 below illustrates. Of course, from the monotonicity just discussed, it is clear that the successive approximation scheme in (8) converges if and only if the two-step operator $T^{(2)}$ has only one fixed point.

We have yet to thoroughly investigate when T has a unique fixed point and when the successive approximation scheme (8) converges. Sufficient conditions for T to be a contraction map on a complete metric space—so that T has a unique fixed point and the successive approximation algorithm in (8) converges to it—are given in Section V, but these conditions are very strong. We make the following conjecture. (It has been proved; see Section VIII.)

Conjecture 1: The reduced-load system of eqs. (7) *always has a unique solution.*

It is easy to see that (7) has a unique solution in the case of two

facilities each with a single server. Extensive numerical testing supports Conjecture 1 when there are only two facilities.

It is, of course, natural to wonder whether the model itself might have multiple equilibrium points, but the exact stochastic process under consideration representing the number of customers from each class in service (with exponential service-time distributions) is an irreducible finite-state, continuous-time Markov chain, which necessarily has a unique equilibrium distribution (Section 2.1 below). Thus, if there are multiple solutions to (7), then they must be an artifact of the approximation.

We now present an example to show that the successive approximation in (8) can fail to converge.

Example 3: (Nonconvergence) To see that the succesive approximation scheme in (8) need not converge to a solution of (7), consider the symmetric model with three facilities, each with one server. Let there be only one customer class, which requires service from all facilities. Let the offered load be α. Then (7) consists of the three equations

$$b^*(1) = B\{1, \alpha[1 - b^*(2)][1 - b^*(3)]\}$$
$$b^*(2) = B\{1, \alpha[1 - b^*(1)][1 - b^*(3)]\}$$
$$b^*(3) = B\{1, \alpha[1 - b^*(1)][1 - b^*(2)]\}$$

in the three unknowns $b^*(1)$, $b^*(2)$, and $b^*(3)$. However, when we apply the operator T, we see that T maps the space of vectors (b_1, b_2, b_3) with $b_1 = b_2 = b_3$ into itself. Since we start with (1, 1, 1) in (8), we only need consider the associated operator $\hat{T}$ on [0, 1], defined by

$$\hat{T}(b) = B[1, \alpha(1 - b)^2] = \frac{\alpha(1 - b)^2}{1 + \alpha(1 - b)^2}.$$

The equation $\hat{T}^{(2)}(b) = b$ leads to the polynomial equation

$$x^5 - x^4 + 2\alpha^{-1}x^3 - 2\alpha^{-1}x^2 + (\alpha + 1)\alpha^{-2}x - \alpha^{-2} = 0$$

for $x = 1 - b$, which factors as

$$(x^2 - x + \alpha^{-1})(x^3 + \alpha^{-1}x - \alpha^{-1}) = 0.$$

The second cubic factor also arises as the solution to $\hat{T}(b) = b$. This cubic polynomial is easily seen to be monotone, so that it has a unique root, which falls in the interval (0, 1). This is the unique symmetric fixed point to the symmetric model. The quadratic term has two roots

$$x = (1 \pm \sqrt{1 - 4\alpha^{-1}})/2,$$

which are real and distinct when $\alpha > 4$. These two roots x_1 and x_2 satisfy $0 \leq x_1 \leq 1$ and $x_1 + x_2 = 1$. The quadratic factor does not have real roots when $\alpha < 4$.

In the case $\alpha = 10$, $\hat{T}$ has unique fixed point $b = 0.607$, which corresponds to the symmetric solution (0.607, 0.607, 0.607). However, $\hat{T}^{(2)}$ has three fixed points: 0.113, 0.607, and 0.887. Hence, the successive approximation scheme (8) fails to converge to the unique symmetric fixed point of T; instead it eventually oscillates between $\mathbf{L} = (0.113, 0.113, 0.113)$ and $\mathbf{U} = (0.887, 0.887, 0.887)$. The exact blocking probability in this case is 0.909, obtained directly from the Erlang loss formula (1). The reduced-load approximation for the customer blocking probability is $b^*(\{1, 2, 3\}) = 1 - (1 - 0.607)^3 = 0.939$. □

Remark: It is significant that with exactly two facilities, the successive approximation scheme in (8) converges if and only if T has a unique fixed point, that is, if and only if (7) has a unique solution. We have already noted that convergence of (8) is equivalent to $T^{(2)}$ having only one fixed point. Obviously, $T^{(2)}$ inherits all fixed points of T, so that if T has multiple fixed points, the (8) will not converge. On the other hand, if (8) fails to converge, then the bounds (L_1, L_2) and (U_1, U_2) obtained from (8) are two distinct fixed points of $T^{(2)}$. In turn, (L_1, U_2) and (U_1, L_2) are two distinct fixed points of T.

This argument extends to certain multifacility models, which include many applications of interest.[25] Suppose that the set of facilities can be partitioned into two subsets such that each customer requires service from one facility in each subset. Let there be n_1 facilities in the first subset, numbered $1 \le i \le n_1$, and n_2 facilities in the other subset, numbered $n_1 + 1 \le i \le n_1 + n_2$. If the successive approximation (8) fails to converge, then $(L_1, \cdots, L_{n_1}, L_{n_1+1}, \cdots, L_{n_1+n_2})$ and $(U_1, \cdots, U_{n_1}, U_{n_1+1}, \cdots, U_{n_1+n_2})$ are distinct fixed points of $T^{(2)}$. It is easy to see that $(L_1, \cdots, L_{n_1}, U_{n_1+1}, \cdots, U_{n_1+n_2})$ and $(U_1, \cdots, U_{n_1}, L_{n_1+1}, \cdots, L_{n_1+n_2})$ are then distinct fixed points of T. □

To summarize the proposed reduced-load approximation, we find approximate blocking probabilities at each facility i by solving (7). To solve (7), we suggest using the successive approximation (8). Successive iterations yield upper and lower bounds on all solutions to (7). If the upper and lower bounds are sufficiently close, then we can stop and use the approximation with some confidence. If the successive approximation bounds are not close, then the whole approach is suspect and we suggest using any solution to (7) with caution. An advantage of solving (7) by (8) is that if (8) converges, then we know there is a unique solution to (7). Moreover, if (8) fails to converge, then we get a warning about the whole approach. Also, (8) is extremely easy to implement. Of course, if (8) fails to converge, then we can look for solutions to (7) by other methods. Alternatively, we might choose to use the final upper bound obtained from (8).

After obtaining the approximate blocking probabilities at each facility, [which usually is a solution to (7), but might not be], we obtain

the approximate total offered loads at each facility via (6) and the approximate blocking probabilities for each class via (5).

Remark 1: To implement the successive approximation in (8) or otherwise solve (7), we need to be able to conveniently calculate the Erlang blocking probability eq. (1) and, for some methods such as Newton's method, its derivatives. For this purpose, we can apply techniques of Jagerman.[27,28] □

Remark 2: The successive approximation in (8) and associated bounds closely parallels a proposed successive approximation algorithm to approximately solve closed networks of queues with a decoupling infinite-server node in Section VI of Ref. 40. The analog in Ref. 40 of the operator T above necessarily has a unique fixed point and the successive approximation scheme also yields bounds on it. However, the successive approximation scheme in Ref. 40 also can fail to converge to the fixed point. Further discussion of the successive approximation in Ref. 40 will appear in a subsequent paper. □

Theorem 2 provides a way to relate the reduced-load approximation (7) to the bound in Corollary 1.1. In particular, we can bound the reduced-load approximation (7) much as we already bounded the exact blocking probability at facility i.

Corollary 2.1: The reduced-load blocking approximation at facility i, that is, any $b^*(i)$ obtained from (7), satisfies

$$B\left\{s_i, \sum_{j\in C_i} \alpha_j \prod_{\substack{k\in A_j \\ k\neq i}} [1 - B(s_k, \hat{\alpha}_k)]\right\} < b^*(i) < B(s_i, \hat{\alpha}_i).$$

Proof: The upper bound is $T^2(\mathbf{1})$ and the lower bound is $T^3(\mathbf{1})$ in the successive approximation (8). □

Let $b^*(A)$ be the reduced-load approximate blocking probability for the subset A obtained by combining (5) and (7). From (5) and Corollary 2.1, we immediately obtain the following bounds for $b^*(A)$.

Corollary 2.2: For each subset A, the reduced-load blocking approximation $b^*(A)$ satisfies

$$1 - \prod_{i\in A}\left(1 - B\left\{s_i, \sum_{j\in C_i} \alpha_j \prod_{\substack{k\in A_j \\ k\neq i}} [1 - B(s_k, \hat{\alpha}_k)]\right\}\right)$$

$$\leq b^*(A) \leq 1 - \prod_{i\in A} [1 - B(s_i, \hat{\alpha}_i)].$$

Note that we have not yet given any lower bounds for the exact blocking probabilities. Obviously, $b(A) \geq \max\{b(i): i\in A\}$, but it seems hard to obtain an improvement. One might conjecture that the exact blocking probability $b(i)$ at facility i is bounded below by the lower

bound in Corollary 2.1, but the following example shows that this is not the case.

Example 4: To see that the lower bound in Corollary 2.1 is not a lower bound on the exact blocking probability, suppose that there are three facilities and two customer classes. Let $s_1 = s_2 = 1$, $s_3 = 3$, $A_1 = \{1, 3\}$, $A_2 = \{2, 3\}$, and $\alpha_1 = \alpha_2 = \alpha$. Then $b(3) = 0$ because at most two of the three servers can be busy at the third facility because of the constraints elsewhere. However, it is easy to see that the lower bound in Corollary 2.1 is strictly positive. □

As a further consequence of Theorem 2, we can show that the bounds in Theorem 1 and its corollaries and the reduced-load approximation in (7) are all asymptotically equivalent as the offered loads per facility become negligible, that is, as $\hat{\alpha}_i \to 0$ for all i.

Corollary 2.3: If $\hat{\alpha}_i \to 0$ *for each* i, *then*

$$\text{(i)}\quad B(s_i, \hat{\alpha}_i)/(\hat{\alpha}_i^{s_i}/s_i!) \to 1,$$

$$\text{(ii)}\quad b^*(i)/B(s_i, \hat{\alpha}_i) \to 1,$$

$$\text{(iii)}\quad \left\{1 - \prod_{i \in A} [1 - B(s_i, \hat{\alpha}_i)]\right\} \Big/ \sum_{i \in A} B(s_i, \hat{\alpha}_i) \to 1$$

$$\text{(iv)}\quad b^*(A) \Big/ \sum_{i \in A} b^*(i) \to 1$$

$$\text{(v)}\quad b^*(A) \Big/ \left\{1 - \prod_{i \in A} [1 - B(s_i, \hat{\alpha}_i)]\right\} \to 1$$

for all subsets A, *where* $b^*(i)$ *and* $b^*(A)$ *are the reduced-load approximations based on* (5) *and* (7).

Proof: Part (i) follows immediately from the form of the Erlang blocking formula in (1). Part (ii) follows from Corollary 2.1 after dividing each term by $B(s_i, \hat{\alpha}_i)$ and letting $\hat{\alpha}_i \to 0$ for all i. To establish the limit for the lower bound, let $\bar{\alpha} = \max\{\hat{\alpha}_i,\ 1 \le i \le n\}$ and $\bar{s} = \min\{s_i,\ 1 \le i \le n\}$. Then

$$\prod_{\substack{k \in A_j \\ k \ne i}} [1 - B(s_k, \hat{\alpha}_k)] \ge [1 - B(\bar{s}, \bar{\alpha})]^{n-1}$$

for all i and j, and $[1 - B(\bar{s}, \bar{\alpha})]^{n-1} \to 1$ as $\hat{\alpha}_i \to 0$ for all i. Hence,

$$B\left\{s_i, \sum_{j \in C_i} \alpha_j \prod_{\substack{k \in A_j \\ k \ne i}} [1 - B(s_k, \hat{\alpha}_k)]\right\} \ge B\{s_i, \hat{\alpha}_i[1 - B(\bar{s}, \bar{\alpha})]^{n-1}\}$$

and

$$B\{s_i, \hat{\alpha}_i[1 - B(\bar{s}, \bar{\alpha})^{n-1}]\}/B(s_i, \hat{\alpha}_i) \to 1$$

as $\hat{\alpha}_i \to 0$ because $B(s_i, \hat{\alpha}_i x)/B(s_i, \hat{\alpha}_i) \to x^{s_i}$ as $\hat{\alpha}_i \to 0$ uniformly in x in any compact subinterval of $(0, \infty)$. Given (i) and (ii), (iii) through (v) are elementary. □

Corollary 2.3 demonstrates that using the more elementary approximations in Theorem 1 and its corollaries instead of the reduced-load approximation (7) is justified if the loads are sufficiently light. Theorem 1 and Corollary 2.3 also suggest that the reduced-load approximation $b^*(A)$ itself might be an upper bound, but the following example shows that the reduced-load approximation $b^*(i)$ obtained from (7) is not an upper bound in general.

Example 5: To see that the reduced-load approximation is not an upper bound, let there be two facilities, each with one server. Let there be two classes with $A_1 = \{1, 2\}$ and $A_2 = \{1\}$, so that $\hat{\alpha}_1 = \alpha_1 + \alpha_2$ and $\hat{\alpha}_2 = \alpha_1$. The reduced-load approximation is determined by the two equations

$$b^*(1) = B\{1, \alpha_1[1 - b^*(2)] + \alpha_2\}$$

$$b^*(2) = B\{1, \alpha_1[1 - b^*(1)]\},$$

from which we easily deduce that $0 < b^*(i) < 1$ for each i, so that $b^*(A_2) = b^*(1) < B(1, \alpha_1 + \alpha_2) = b(1)$. □

Remark: One might conjecture that all the approximations for the exact blocking probability $b(i)$ are asymptotically correct as the loads decrease, but Example 2 shows that this is not nearly the case. If $\alpha_2 = 100\alpha_1$ there, then $b(1)/\alpha_1^{10} \to 101$, while $B(s_1, \hat{\alpha}_1)/\alpha_1^{10} \to (101)^{10}$ as $\alpha_1 \to 0$. However, a positive result for large symmetric models appears in Corollary 3.2 below. □

1.6 Symmetric solutions to symmetric models

In this section we consider the special case of symmetric models in which all facilities have s servers and offered load $\hat{\alpha}$, and all classes require service from m facilities. To have full model symmetry, we also assume that there is a class requiring service from each subset of m facilities, and that the offered loads are the same for each class. We also assume that the arrival rates and service rates are the same for all classes.

If we restrict attention to symmetric solutions to symmetric models, then the reduced-load system of eqs. (7) simplifies to the single polynomial equation in one variable

$$b^* = B(s, \hat{\alpha}(1 - b^*)^{m-1}), \tag{9}$$

where $b^*(i) = b^*$ for all i. Since the right side of (9) is continuous and decreasing as a function of b^*, (9) has a unique solution, which is easy to find.

Note that we have restricted attention to symmetric solutions of (7) in order to obtain the single eq. (9). We have not yet ruled out asymmetric solutions to symmetric models. However, we conjecture that none exist. (See Section VIII.)

Conjecture 2 (Corollary to Conjecture 1): The symmetric solution [that is, the solution to (9)*] to the reduced-load approximation eqs.* (7) *is the only solution for a symmetric model.*

To investigate the accuracy of the approximation (5) through (9), we investigate the asymptotic behavior of symmetric models as $n \to \infty$ with the offered load per facility, $\hat{\alpha}$, and the number of facilities required per class, m, held fixed. In Section III we prove that the approximation (5) through (9) is asymptotically correct as $n \to \infty$ under these conditions. Note that since we fix the offered load per facility, $\hat{\alpha}$, this limit does not correspond to light traffic.

To state the main result, let Y_{ni} be the number of busy servers at facility i and let Z_{nk} be the proportion of the facilities with k busy servers (both in steady state) when there are n facilities. Let $\xrightarrow{p}$ denote convergence in probability.

Theorem 3: If $n \to \infty$ with $\hat{\alpha}$ and m held fixed for the symmetric model, then

(a) $Z_{nk} \xrightarrow{p} \beta_k$ as $n \to \infty$ for each k, where β_k satisfies the M/G/s/loss formula

$$\beta_k = (\xi^k/k!) \Big/ \sum_{l=0}^{s} (\xi^l/l!) \tag{10}$$

with

$$\xi = \hat{\alpha}(1 - \beta_s)^{m-1}; \tag{11}$$

that is, β_s is the unique symmetric solution to (9).

(b) For any finite subset H, the random variables Y_{ni}, $i \in H$, are asymptotically mutually independent as $n \to \infty$.

We establish Theorem 3 in Section III by first focusing on the stochastic process representing the proportion of facilities with k busy servers at time t, $1 \le k \le s$ and $t \ge 0$. The key result is a functional law of large numbers for this sequence of stochastic processes as $n \to \infty$ (Theorem 8). The analysis is challenging because this stochastic process is not Markov.

From Theorem 3, we easily obtain our desired corollary.

Corollary 3.1: For symmetric models, the symmetric reduced-load approximation in (5) *through* (9) *is asymptotically correct as $n \to \infty$ with $\hat{\alpha}$ and m held fixed.*

We can combine Corollaries 2.3 and 3.1 to conclude that the bounds in Theorem 1 and its corollaries are also asymptotically correct with light loads.

Corollary 3.2: For symmetric models, the bounds in Theorem 1 *and its corollaries are asymptotically correct as* $n \to \infty$ *and* $\hat{\alpha} \to 0$; *that is, for each integer* k *and each positive* ϵ, *there is a critical offered load* α_0 *and an integer-valued function* $n(\alpha)$ *such that*

$$\left| \frac{b(A)}{kB(s, \hat{\alpha})} - 1 \right| < \epsilon,$$

for all $\hat{\alpha} < \alpha_0$ *and* $n \geq n(\hat{\alpha})$, *where* A *is a subset with* k *facilities.*

Example 5 shows that the reduced-load approximation is not an upper bound on the actual blocking probability in general. Example 1 shows that the symmetric reduced-load approximation $b^*(i)$ for the blocking probability at each facility in a symmetric model need not be an upper bound either. However, we make the following conjecture.

Conjecture 3: The reduced-load approximation for the blocking probability of each customer in a symmetric model in which the number of facilities per customer is fixed, obtained by combining (5) *and* (9), *is always an upper bound.*

Remark 1: In our symmetric model each customer requires service from m facilities. Instead, as in Ref. 21, we could have different types of customers, with customers of type m requiring service from m facilities. The facilities remain symmetric with this change, so that if we still restrict attention to symmetric solutions to the symmetric model, then we again obtain a single polynomial equation in one variable. In particular, suppose that we have M types, numbered from 1 to M. If we let $\tilde{\alpha}_m$ be the total offered load of type m at each facility, then we obtain (9) with the second argument of B replaced by $\sum_{m=1}^{M} \tilde{\alpha}_m (1 - b^*)^{m-1}$. Consequently, it is easy to approximately solve the models in Ref. 21 and generalizations in which each facility has s servers. With this extended symmetric model we abandon Conjecture 3. It is easy to get a counterexample by modifying Example 1 to introduce additional customers that require service from only one facility and have negligible offered load. □

Remark 2: The reduced-load approximation has the potential of being a powerful and flexible approximation tool if we judiciously control the amount of symmetry. For example, we can obtain a richer class of database locking models by requiring only partial symmetry. Some regions of the database may be requested much more than others. There may also be a tendency for the items requested in a given transaction to cluster together. These general features can be represented by partitioning the database into mutually exclusive subsets and assuming symmetry only within each subset. In addition, we can introduce various types of transactions, as in Remark 1 above. The partial symmetry causes the reduced-load approximation to be a

system of k equations in k unknowns, where k is the number of subsets in the partition. The number of transaction types does not increase the number of equations. Again, the successive approximation (8) can be applied. □

Remark 3: Mitra and Weinberger established Corollary 3.2 for multiple-customer types in the special case of one server per facility via their asymptotic analysis.[21] Heyman also has a different proof of Corollary 3.2 in the special case of one server per facility, assuming that the total offered load in the network is fixed as $n \to \infty$.[23] □

1.7 A few numerical examples

Table I compares the approximations in Theorem 1 and its corollaries with the reduced-load approximation in (5) through (9) for several symmetric models. The various approximations were calculated "by hand" at the terminal using the Erlang blocking formula algorithms of Jagerman[28] (coded by Moshe Segal). The approximations are all independent of the number of facilities, so n is not specified. Based on Theorem 3, the reduced-load approximation in (9) is asymptotically correct for large n. The offered load per facility $\hat{\alpha}$ in (3) is chosen so that the nominal blocking per facility (the bound in Corollary 1.1) has a specified value: 0.10 in the first six cases, 0.02 in the next three cases, and 0.01 in the last three cases.

From Table I (and intuition), it is apparent that the quality of the bounds as approximations is a decreasing function of the number s of servers per facility, the offered load per facility $\hat{\alpha}$, and the number m of facilities per class. In the case of nominal blocking per facility of

Table I—The approximate blocking probability for each customer class in symmetric models: a comparison of the approximation procedures

Servers per Facility s	Offered Load per Facility $\hat{\alpha}$	Facilities per Class m	Summation Bound in Corollary 1.2	Product Bound in Theorem 1	Reduced-Load Approximation (9)
1	0.11111	2	0.200	0.190	0.175
10	7.51	2	0.200	0.190	0.146
50	49.6	2	0.200	0.190	0.126
1	0.11111	3	0.300	0.271	0.234
10	7.51	3	0.300	0.271	0.178
50	49.6	3	0.300	0.271	0.157
1	0.0204	5	0.100	0.096	0.089
10	5.087	5	0.100	0.096	0.072
50	40.27	5	0.100	0.096	0.057
1	0.010101	2	0.0200	0.0199	0.0197
10	4.464	2	0.0200	0.0199	0.0192
50	37.90	2	0.0200	0.0199	0.0180

Table II—Examples of the successive approximations in (8) for the reduced-load approximation with the symmetric model in the first three cases of Table I

Servers per facility s	1	10	50
Offered load per facility $\hat{\alpha}$	0.11111	7.51	49.6
Facilities per class m	2	2	2
Bound in Corollary 1.2	0.200	0.200	0.200
Bound in Theorem 1	0.190	0.190	0.190
Bound in Corollary 1.1	0.100	0.100	0.100
Iteration one	0.089	0.069	0.051
Iteration two	0.0919	0.078	0.074
Iteration three	0.0916	0.076	0.063
Iteration four	0.0917	0.077	0.068
Reduced-load approx. (9)	0.0917	0.076	0.065
Approximate blocking for each class by (5) and (9)	0.175	0.146	0.126

0.01 and only two facilities required per class (the last three cases), the simple summation bound in Corollary 1.2 seems to be adequate. However, the case of $s = 50$ and $m = 5$ produces perhaps a surprisingly large discrepancy between (9) and the bounds.

Table II displays the outcomes of the successive approximations in (8) applied to the first three cases in Table I. The successive iterations describe the blocking per facility, as in (7) and (9). Then (5) is applied to obtain the blocking per class. From Table II it is apparent that about five iterations yields adequate accuracy, that is, getting close enough to the fixed point (9). In these examples the successive approximation scheme in (8) converges to the unique symmetric fixed point of (9).

Table III compares the approximations with exact blocking probabilities for different numbers of facilities in the special case of a symmetric model with $s = 1$ (one server per facility) and ($m = 2$) (two facilities required per class). When $s = 1$, the exact blocking probability is relatively easy to compute because, with exponential service times having mean one (which we can assume without loss of generality by Theorem 4 and Corollary 4.2), the number of customers in service (which is the number of busy servers divided by m) is a birth-and-death process with death rate $\mu(k) = k$ and birth rate

$$\lambda(k) = (n\hat{\alpha})\left(\frac{(n - km)(n - km - 1)\cdots(n - km - m + 1)}{n(n - 1)\cdots(n - m + 1)}\right). \quad (12)$$

The data in Table III for this special case were obtained from D. P. Heyman (personal communication). This case is consistent with Theorem 3, which establishes that (9) is asymptotically correct as $n \to \infty$. Table III leads us to conjecture that the exact blocking probability for each class is increasing in n in this case. More generally, we make the following conjecture.

Table III—Comparison of approximations with exact blocking probabilities for symmetric models when $s = 1$ (one server per facility) and $m = 2$ (two facilities required per class)

Number of Facilities n	Offered Load per Facility $\hat{\alpha}$	Exact Blocking Probability	Reduced-Load Approximation (9)	Product Bound in Theorem 1	Summation Bound in Corollary 1.2
2	0.010101	0.0100	0.0197	0.0199	0.0200
4		0.0165	0.0197	0.0199	0.0200
8		0.0183	0.0197	0.0199	0.0200
40		0.0195	0.0197	0.0199	0.0200
100		0.0196	0.0197	0.0199	0.0200
2	0.111111	0.100	0.175	0.190	0.200
4		0.154	0.175	0.190	0.200
8		0.168	0.175	0.190	0.200
40		0.1735	0.175	0.190	0.200
100		0.1744	0.175	0.190	0.200
2	1.0000	0.500	0.618	0.750	1.000
4		0.600	0.618	0.750	1.000
8		0.611	0.618	0.750	1.000
40		0.6168	0.618	0.750	1.000
100		0.6176	0.618	0.750	1.000

Conjecture 4: The exact blocking probability for each customer class in a symmetric model is a nondecreasing function of the number n of facilities when the offered load per facility $\hat{\alpha}$ and the number m of facilities per customer are held fixed.

Remark: Conjecture 3 is a corollary to Conjecture 4 and Theorem 3. □

For typical blocking probabilities (0.001 through 0.2), the quality of the approximations appears to be a decreasing function of the offered load per facility (or nominal blocking probability), but this is evidently not true over the full range. The middle four cases in Table III provide greater relative differences than the last four cases, comparing (9) with the exact values.

As our final example in this subsection, we consider a communication network with traffic from several different sources to a common destination, as depicted in Fig. 1. Traffic from each source needs two lines: one line in a facility associated with that source plus one line in a final facility shared by all sources. When there are n sources, there are n customer classes and $n + 1$ facilities. For each i, $1 \le i \le n$, class i requires one server from facility i and one server from facility $n + 1$. Note that this example has the special structure mentioned in the remark following Example 3, so that for the reduced-load approximation the successive approximation scheme in (8) converges if and only if the operator T has a unique fixed point.

Tables IV and V give numerical results obtained by J. T. Wittbold[25] for several cases in which n equals 2 and 3, respectively. We display

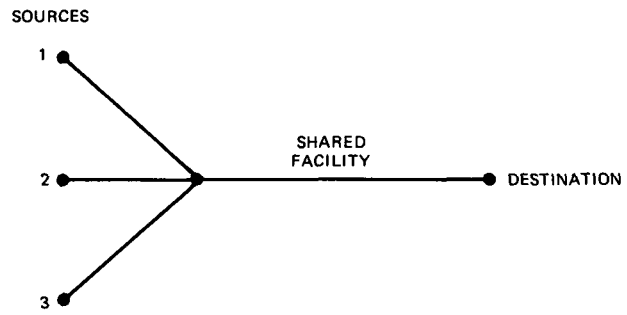

Fig. 1—A communication network with four facilities and three customer classes.

the exact blocking probability and the reduced-load approximation for each facility and for each customer class. The successive approximation converged quickly in every case. For the customer classes, we also display the product bounds and the approximation obtained by taking the product of the exact facility nonblocking probabilities (the last column). This last column helps assess how much of the error is due to assuming facility independence.

For the cases with high blocking probabilities, for example, Case 1 in Tables IV and V, the reduced-load approximation is much better than the product bound, as expected. Overall, the reduced-load approximation appears adequate for engineering purposes. For lighter loads, for example, Cases 4 through 6 in Table IV and V, the product bound seems adequate for most engineering purposes. It should be effective for properly sizing facilities given forecasting data.

1.8 Existence, uniqueness, and insensitivity

It is significant that we have assumed nothing about the service-time distributions except that they have finite means. For applications, experience indicates that call attempts can often be modeled reasonably by a Poisson process, but that virtual circuit holding-time distributions are often not nearly exponential.[41,43] In Section IV we rigorously establish that a steady-state blocking probability exists, is unique, and depends on the service-time distributions only through their means. For this, we apply the theory of Generalized Semi-Markov Processes (GSMPs) and the associated theory of insensitivity.[44–46]

It turns out that the model we consider also can be regarded as a special case of a model analyzed by Kaufman[7] of blocking in a single facility in which customers request several servers and there is a general resource-sharing policy. The connection to Kaufman's single-facility model is made by simply combining our n facilities and implementing our sharing scheme as one of his general sharing policies. The insensitivity property and the exact formula for the blocking are

Table IV—Comparison of approximations with exact blocking probabilities for the communication network example in Fig. 1 with $n = 2$ sources

Case Number	Facility Number	Number of Servers s_i	Offered Load α_i	Blocking Probability at Facility i		Blocking Probability for Customer Class i			
				Reduced Load	Exact	Product Bound	Reduced Load	Exact	Product of Exact Probabilities
1	1	20	30	0.209	0.199	0.61	0.42	0.41	0.43
	2	15	15	0.059	0.025	0.48	0.31	0.30	0.31
	$n+1$	30	—	0.266	0.293	—	—	—	—
2	1	30	30	0.067	0.041	0.29	0.19	0.17	0.19
	2	15	15	0.116	0.101	0.33	0.23	0.23	0.24
	$n+1$	40	—	0.133	0.151	—	—	—	—
3	1	30	30	0.006	0.000	0.45	0.36	0.36	0.36
	2	15	15	0.029	0.017	0.48	0.38	0.37	0.38
	$n+1$	30	—	0.360	0.365	—	—	—	—
4	1	40	30	0.008	0.002	0.068	0.055	0.050	0.051
	2	20	15	0.034	0.030	0.097	0.080	0.075	0.077
	$n+1$	50	—	0.048	0.049	—	—	—	—
5	1	42	30	0.006	0.006	0.023	0.021	0.017	0.017
	2	23	15	0.011	0.012	0.029	0.026	0.023	0.024
	$n+1$	56	—	0.014	0.015	—	—	—	—
6	1	42	27	0.0017	0.0012	0.005	0.005	0.004	0.004
	2	23	13	0.0036	0.0033	0.007	0.007	0.006	0.006
	$n+1$	56	—	0.0027	0.0027	—	—	—	—

Table V—Comparison of approximations with exact blocking probabilities for the communication network example in Fig. 1 with $n = 3$ sources

Case Number	Facility Number	Number of Servers s_i	Offered Load α_i	Blocking Probability at Facility i		Blocking Probability for Customer Class i			
				Reduced Load	Exact	Product Bound	Reduced Load	Exact	Product of Exact Probabilities
1	1	40	30	0.001	0.000	0.30	0.19	0.19	0.19
	2	10	20	0.446	0.452	0.67	0.55	0.55	0.56
	3	20	17	0.027	0.018	0.35	0.21	0.20	0.21
	$n+1$	50	—	0.190	0.192	—	—	—	—
2	1	40	30	0.008	0.003	0.16	0.06	0.05	0.05
	2	10	20	0.523	0.523	0.61	0.54	0.54	0.54
	3	20	17	0.067	0.063	0.22	0.11	0.10	0.11
	$n+1$	61	—	0.044	0.044	—	—	—	—
3	1	40	30	0.012	0.009	0.05	0.03	0.02	0.02
	2	10	8	0.115	0.116	0.15	0.13	0.13	0.13
	3	20	17	0.078	0.079	0.12	0.10	0.09	0.09
	$n+1$	63	—	0.020	0.016	—	—	—	—
4	1	40	30	0.013	0.013	0.03	0.02	0.02	0.02
	2	10	8	0.119	0.121	0.13	0.13	0.12	0.12
	3	20	17	0.083	0.085	0.10	0.09	0.09	0.09
	$n+1$	67	—	0.007	0.003	—	—	—	—
5	1	40	30	0.013	0.011	0.028	0.024	0.019	0.020
	2	10	5	0.017	0.017	0.031	0.028	0.025	0.026
	3	20	13	0.017	0.016	0.031	0.028	0.024	0.025
	$n+1$	60	—	0.011	0.009	—	—	—	—
6	1	40	28	0.0059	0.0045	0.016	0.014	0.011	0.012
	2	10	4	0.0050	0.0048	0.015	0.014	0.012	0.012
	3	20	12	0.0091	0.0084	0.019	0.018	0.016	0.016
	$n+1$	57	—	0.0085	0.0075	—	—	—	—

thus available from Ref. 7. (Reference 7 also mentions other related work.) We contribute to Ref. 7 by verifying the conjecture on p. 1477 there that the insensitivity property holds for arbitrary service-time distributions, not just service-time distributions with rational Laplace-Stieltjes transforms. (The insensitivity analysis for our model extends to the setting of Ref. 7, but the bounds and approximations do not.)

Insensitivity properties in queueing have a long history, going all the way back to Erlang.[47] Insensitivity theory for queueing networks is largely due to Baskett, Chandy, Muntz and Palacios[48] and Kelly.[49] It is now understood[50] that this theory can be viewed as a consequence of the earlier work by Matthes[51] on "bedienungsprozesse" or GSMPs.

As Kaufman observes,[7] his model is equivalent to a closed multiclass BCMP network[48] with the addition of extra population constraints. Without the population constraints, we could simply apply the insensitivity theory developed by Baskett et al.[48] and Kelly,[49] which was extended to arbitrary service-time distributions by Barbour;[52] for example, we could apply Section 3.3 of Ref. 6). As observed by Lam,[53] it is possible to extend the insensitivity theory to closed networks with population constraints, but it is perhaps more appropriate to recognize that the closed network, with or without population constraints, is a GSMP, and the insensitivity theory for GSMPs can be applied directly. The direct approach via GSMPs is contained in Burman et al.[8] The analysis in both Kaufman[7] and Burman et al.[8] requires the addition of Ref. 46 to treat arbitrary service-time distributions. The technical details here for establishing existence, uniqueness, and insensitivity appear in Section IV.

1.9 Non-Poisson arrival processes

We now indicate how the reduced-load approximation (5) through (7) can be combined with previous approximations for the blocking in a single facility with non-Poisson arrival processes to generate approximations for blocking probabilities in the multifacility model here when we relax the assumption that the arrival process of each class is a Poisson process.

We assume that the arrival process of each class is a general stationary point process[54] partially characterized by its arrival rate λ_j and peakedness z_j. (The facilities are thus G/GI/s/loss systems instead of M/GI/s/loss systems. See Refs. 55 through 57 and references in these sources for background on peakedness.) As before, we assume that the arrival processes of the different classes and all the service times are mutually independent.

We regard the arrival process at facility i as the superposition of the arrival processes of those classes requiring service from facility i.

Hence, paralleling (3), we define the peakedness of the arrival process at facility i as

$$\hat{z}_i = \sum_{j \in C_i} \alpha_j z_j / \hat{a}_i, \tag{13}$$

where α_j is the offered load and z_j is the peakedness for class j. Formula (13) is the standard peakedness approximation for a superposition process, but it is based on the assumption that the service rates are the same for all classes, which is not necessarily the case here. Since we do not account for this difficulty, (13) should perform better if the service rates μ_j do not vary much. References 55 through 57 describe ways to determine the peakedness z_j for each class; one relatively simple way is the heavy-traffic approximation in (4) of Ref. 57, but other more involved methods are usually more accurate.

For our new reduced-load approximation, we again use (5) and (6). We propose Hayward's approximation to extend (7).[55-57] However, with the non-Poisson arrival processes we must first carefully distinguish different notions of blocking. Let $b_c(A)$, $b_{Ci}(A)$, and $b_T(A)$ be the probability that all servers are busy in at least one facility in the set A at the instant of an arbitrary arrival, an arrival to facility i, and at an arbitrary time, respectively (the overall call congestion, the facility-i call congestion and the time congestion). Let b_j and $b_j(i)$ be the blocking probability for class j overall and at facility i, respectively. We are primarily interested in b_j, $b_{Ci}(i)$, and $b_T(A)$.

We apply Hayward's approximation to approximate $b_{Ci}(i)$ as if facility i were in isolation. We use the peakedness $\hat{z}_i$ in (13) to modify (7) in the usual way:

$$b_{Ci}(i) = B(s_i/\hat{z}_i,\ \bar{\alpha}_i/\hat{z}_i), \tag{14}$$

where $B(s, \alpha)$ is the Erlang blocking formula in (1) extended to noninteger s, as described in Refs. 27 and 28, and instead of (6) $\bar{\alpha}_i$ is

$$\bar{\alpha}_i = \sum_{j \in C_i} \alpha_j \prod_{\substack{k \in A_j \\ k \neq i}} [1 - b_T(k)]. \tag{15}$$

In (15) we use $b_T(k)$ to approximately represent the blocking probability at facility k seen by an arbitrary arrival to facility i. This involves an aspect of the basic facility independence approximation in (5).

We obtain the approximate time congestion for facility i by using the approximation

$$b_T(i) = b_{Ci}(i)/\hat{z}_i \tag{16}$$

(see page 695 of Ref. 57). [A significant improvement should be possible by using (16) of Ref. 56 with the equivalent random method instead of (16) above.] Hence, instead of (7), we obtain the following

system of n equations in the n unknowns $b_{Ci}(i)$ by combining (14) through (16):

$$b_{Ci}(i) = B\left(s_i/\hat{z}_i,\ (1/\hat{z}_i)\sum_{j\in C_i}\alpha_j \prod_{\substack{k\in A_j\\ k\neq i}}\{1 - [b_{Ck}(k)/\hat{z}_k]\}\right). \tag{17}$$

Since $\hat{z}_i$ are fixed positive scalars in (17), the successive approximation in (8) applies here as well; that is, Theorem 2 and Corollaries 2.1 and 2.2 extend easily.

Given that we have obtained $b_{Ci}(i)$ and $b_T(i)$ via (16) and (17), we combine (5) and (16) to obtain the time congestion for an arbitrary subset A, that is,

$$b_T(A) = 1 - \prod_{i\in A}[1 - b_T(i)] = 1 - \prod_{i\in A}\{1 - [b_{Ci}(i)/\hat{z}_i]\}. \tag{18}$$

Next we obtain the blocking for class j at facility i by combining our approximations with Fredericks' approximation for parcel blocking, (23) in Ref. 56. We obtain

$$b_j(i) = b_T(i) + \frac{(z_j - 1)}{(\hat{z}_i - 1)}[b_{Ci}(i) - b_T(i)]. \tag{19}$$

Finally, we obtain the overall blocking for class j by combining (5) and (19), that is,

$$b_j = 1 - \prod_{i\in A_j}[1 - b_j(i)]. \tag{20}$$

The approximations for $b_{Ci}(i)$, $b_T(A)$, and b_j in (17), (18), and (20) have yet to be tested, but experience with the individual approximation steps suggest that the combined procedure is promising.

II. THE BOUNDS

2.1 The exact blocking formula

As a basis for proving Theorem 1, we first calculate the exact blocking probabilities $b(A)$. For this purpose, let N_j represent the steady-state number of class j customers in service. The distribution of the vector $(N_1, \cdots, N_c)$ is conveniently described in terms of the random vector $(N_1^\infty, \cdots, N_c^\infty)$, where N_j^∞ represents the steady-state number of class j customers in service when all n facilities have infinitely many servers, but otherwise the model is the same. Of course, in the infinite-server model the steady-state distribution is easy to describe because there is no blocking, so that there is no interaction among the classes; that is, the random variables $N_1^\infty, \cdots, N_c^\infty$ are independent. From basic results for the M/G/∞ congestion model,[6] the steady-state distribution is

$$P(N_j^\infty = k_j,\ 1 \le j \le c) = \prod_{j=1}^{c} P(N_j^\infty = k_j) = \prod_{j=1}^{c} \left(\frac{\alpha_j^{k_j}}{k_j!}\right). \qquad (21)$$

As in closed Jackson networks of queues, the steady-state distribution of $(N_1, \cdots, N_c)$ is obtained from (21) by simply conditioning. (See Section 1.6 of Ref. 6.) We defer the proof until Section IV.

Theorem 4: The steady-state distribution of $(N_1, \cdots, N_c)$ exists, is unique, depends on the service-time distributions only through their means, and has the form

$$P(N_j = k_j,\ 1 \le j \le c)$$

$$= P\left(N_j^\infty = k_j,\ 1 \le j \le c \;\middle|\; \sum_{j \in C_i} N_j^\infty \le s_i,\ 1 \le i \le n\right)$$

$$= \frac{P(N_j^\infty = k_j,\ 1 \le j \le c)}{P\left(\sum_{j \in C_i} N_j^\infty \le s_i,\ 1 \le i \le n\right)}.$$

Of course, Theorem 4 can be used to give an exact expression for the blocking probability $b(A)$. Let Y_i represent the number of busy servers at facility i in our model, that is, $Y_i = \sum_{j \in C_i} N_j$.

Corollary 4.1: For each subset A, $b(A) = 1 - P(Y_i < s_i,\ i \in A)$.

However, we apply Theorem 4 only via the following elementary consequence.

Corollary 4.2: The distribution of $(N_1^\infty, \cdots, N_c^\infty)$ and thus also the distributions of $(N_1, \cdots, N_c)$ and $(Y_1, \cdots, Y_n)$ depend on the vectors of arrival rates $(\lambda_1, \cdots, \lambda_c)$ and service rates $(\mu_1, \cdots, \mu_c)$ only through the vector of offered loads $(\alpha_1, \cdots, \alpha_c)$, where $\alpha_j = \lambda_j/\mu_j$.

Remark: It is significant in Corollary 4.2 that there is not just one degree of freedom, corresponding to the choice of our measuring unit, but c degrees of freedom. For example, we can arbitrarily select the service rate μ_j for each class j, as long as the offered load α_j is as originally specified. In fact, for us it will be convenient to make all service rates identical. (See the proofs of Theorems 5 and 7.)

2.2 Proof of Corollary 1.1

To give a direct proof of Corollary 1.1, we establish a stronger stochastic comparison. Let $N(s, \alpha)$ represent the steady-state number of busy servers in an M/G/s/loss system with s servers and offered load α. We use the notion of Monotone-Likelihood-Ratio (MLR) ordering.[30] An integer-valued random variable X_1 is said to be less than or equal to another integer-valued random variable X_2 in the MLR ordering, denoted by $X_1 \le_r X_2$, if

$$\frac{P(X_1 = k+1)}{P(X_1 = k)} \leq \frac{P(X_2 = k+1)}{P(X_2 = k)} \tag{22}$$

for all k. (We also require that the supports be ordered intervals; that is, $P(X_i = k) > 0$ for integers $k \in [a_i, b_i]$, where $-\infty \leq a_i < b_i \leq +\infty$, $a_1 \leq a_2$, and $b_1 \leq b_2$.) The MLR ordering is useful largely because it implies ordinary *stochastic order*, namely,

$$Ef(X_1) \leq Ef(X_2) \tag{23}$$

for all nondecreasing functions f for which the expectations are well defined.[30,35]

Theorem 5: For each facility i, $Y_i \leq_r N(s_i, \hat{\alpha}_i)$.

Proof: First, make all service-time distributions exponential with mean one, without altering any of the offered loads. By Theorem 4 and Corollary 4.2, this does not alter the steady-state distribution of $(N_1, \cdots, N_c)$. Next apply Theorem 5 of Ref. 30. The service rate in both systems is k when there are k busy servers. The arrival rate at facility i in the actual system is always less than or equal to $\hat{\alpha}_i$. It is less when there is blocking elsewhere. Of course, the support of both random variables is the set $\{0, 1, \cdots, s_i\}$. □

Proof of Corollary 1.1: Apply Theorem 5 and (23), noting that

$$b(i) = P(Y_i \geq s_i) \leq P[N(s_i, \hat{\alpha}_i) \geq s_i] = B(s_i, \hat{\alpha}_i). \quad \square \tag{24}$$

Having proved Corollary 1.1, we immediately obtain Corollary 1.2 by virtue of the Bonferroni inequalities (see page 110 of Ref. 29).

2.3 Plausible stochastic comparisons

It is natural to conjecture that Corollary 1.2 could be improved to Theorem 1 by exploiting the exact relationship in Corollary 4.2 and establishing the inequality (4) or, equivalently, that

$$P(Y_i < s_i, \, i \in A) \geq \prod_{i \in A} P(Y_i < s_i). \tag{25}$$

Formula (25) would follow from the random variables Y_i, $i \in A$, being associated or just positively quadrant dependent (see pages 29 and 142 of Ref. 58). Unfortunately, however, (25) is not true in general, as we show in Example 6 below.

One might also try to establish Theorem 1 via certain multivariate stochastic comparisons. In particular, it is natural to consider the multivariate versions of the MLR ordering $\leq_r$ and the stochastic ordering $\leq_{st}$ defined in (22) and (23) (see Refs. 59 and 60). The extension of $\leq_{st}$ is defined again by (23). It is natural to conjecture that

$$(Y_1, \cdots, Y_n) \leq_{st} [N_1(s_1, \hat{\alpha}_1), \cdots, N_n(s_n, \hat{\alpha}_n)], \tag{26}$$

where the variables $N_i(s_i, \hat{\alpha}_i)$ are mutually independent. It is also natural to conjecture the weaker relationships

$$P(Y_i \geq k_i,\ 1 \leq i \leq n) \leq \prod_{i=1}^{n} P[N(s_i, \hat{\alpha}_i) \geq k_i] \tag{27}$$

and

$$P(Y_i \leq k_i,\ 1 \leq i \leq n) \geq \prod_{i=1}^{n} P[N(s_i, \hat{\alpha}_i) \leq k_i] \tag{28}$$

for all n-tuples $(k_1, \cdots, k_n)$. However, in Example 6 below we show that (27) is not valid, which implies that (26) and the stronger ordering with $\leq_r$ instead of $\leq_{st}$ in (26) are not valid either. However, it turns out that (28) is valid, and that is the key to establishing Theorem 1.

Example 6: To see that (25) and (27) need not hold, consider the symmetric model with $n = c = 3$, $s_1 = s_2 = s_3 = 1$, $A_1 = \{1, 2\}$, $A_2 = \{1, 3\}$, $A_3 = \{2, 3\}$, $\mu_1 = \mu_2 = \mu_3 = 1$, and $\lambda_1 = \lambda_2 = \lambda_3 = \alpha$. Then $b(A_j) = 3\alpha/(1 + 3\alpha)$ for all classes j and $b(i) = 2\alpha/(1 + 3\alpha)$ for all facilities i. Hence, for $\alpha > 1$

$$1 - b(A_1) = \frac{1}{1 + 3\alpha} < \left(\frac{1 + \alpha}{1 + 3\alpha}\right)^2 = [1 - b(1)][1 - b(2)], \tag{29}$$

so that (25) fails. On the other hand,

$$1 - b(A_1) = \frac{1}{1 + 3\alpha} > \left(\frac{1}{1 + 2\alpha}\right)^2 = [1 - B(s_1, \hat{\alpha}_1)]^2 \tag{30}$$

so that the conclusion of Theorem 1 still holds in this case.

To see that (27) can fail too, let $(k_1, k_2, k_3) = (1, 1, 0)$. Then

$$P(Y_i \geq k_i, \quad 1 \leq i \leq 3) = \alpha/(1 + 3\alpha), \tag{31}$$

while

$$\prod_{i=1}^{3} P(N(s_i, \hat{\alpha}_i) \geq k_i) = [2\alpha/(1 + 2\alpha)]^2. \tag{32}$$

Hence, for $\alpha^2 < 1/8$, (27) fails. On the other hand, it is easy to see that (28) does still hold in this example. By symmetry, it suffices to consider only the two triples (1, 1, 0) and (1, 0, 0). □

In summary, Example 6 shows that none of the plausible relations (4), (25), (26), and (27) is valid, but the validity of Theorem 1 and (28), which would imply Theorem 1, remains open. We now proceed to establish (28).

2.4 Proof of Theorem 1

To prove Theorem 1 we establish (28). To establish (28), we develop

a certain multivariate variation of Theorem 5 in Ref. 30. In particular, we develop a general stochastic comparison result for continuous-time non-Markov jump processes in which the intensities of moving into certain sets are always greater for one process than the other. The results here are a special case of the general theory developed in Ref. 31. They also can be obtained from the related work of Massey.[36-38]

For our comparison result, we consider an arbitrary finite state space S. (It will be clear that similar results hold for infinite state spaces, but it suffices for us to consider a finite state space.) Let the space $\mathscr{P} \equiv \mathscr{P}(S)$ of all probability measures P on S be endowed with an order relation $\leq$ defined by $P_1 \leq P_2$ if $P_1(A) \leq P_2(A)$ for all subsets A of S in some class $\mathscr{A}$. (The order relation $\leq$ is obviously reflexive and transitive, but it is not necessarily a partial order because it need not be antisymmetric: $P_1 \leq P_2$ and $P_2 \leq P_1$ together do not necessarily imply that $P_1 = P_2$; the relation will be a partial order if $\mathscr{A}$ is a determining class.[61]) Since S is finite, the order relation is closed, that is, it is preserved under limits: If $P_{1n} \leq P_{2n}$ in $(\mathscr{P}, \leq)$ for all n, $P_{in}(\{s\}) \to P_i(\{s\})$ as $n \to \infty$ for each i and $s \in S$, then $P_1 \leq P_2$. [In our application S will be a finite subset of R^n, but $\leq$ will not correspond to ordinary stochastic order on $\mathscr{P}(S)$ as defined in (23).]

The first process $Y_1(t)$ will be a Continuous-Time Markov Chain (CTMC) with infinitesimal transition rates (generator) $q_1(s; A)$, defined as usual for $s \in S$ and $A \subseteq S$ in terms of its transition function by

$$P(Y_1(t + h) \in A \mid Y_1(t) = s) = hq_1(s; A) + o(h), \qquad (33)$$

for $s \notin A$, where $o(h)$ represents a quantity that converges to zero after dividing by h.

The second process $Y_2(t)$ will also be a continuous-time jump process with the jumps governed by infinitesimal transition rates, but as in Ref. 30 these rates may depend on additional information other than the current state, such as the history of the process. Let the additional information at time t be $\Gamma(t)$, and let γ represent a possible value. [In our application the process $Y_2(t)$ represents the number of busy servers at each facility, and the additional information $\Gamma(t)$ is the number of customers of each class in service.] We assume that the process $[Y_2(t), \Gamma(t)]$ is a CTMC on the product state space $S \times S'$, where S' as well as S is finite. Let $q_2(s, \gamma; A)$ be the transition function for $[Y_2(t), \Gamma(t)]$, defined by

$$P([Y_2(t), \Gamma(t)] \in A \mid Y_2(t) = s, \Gamma(t) = \gamma) = hq_2(s, \gamma; A) + o(h) \quad (34)$$

for $(s, \gamma) \notin A$ and $A \subseteq S \times S'$. We shall also use the transition function for $Y_2(t)$, defined by $q_2(s, \gamma; A \times S')$ for $A \subseteq S$ and $s \notin S$.

Just as in Ref. 30, the idea here is to compare the processes $Y_1(t)$

and $Y_2(t)$ by comparing the transition intensities in the space S, requiring that the comparisons hold uniformly in the extra information $\Gamma(t)$, which must be added to $Y_2(t)$ to make $Y_2(t)$ Markov. Of course, a major complication here is the multidimensional state space S. Following Kester[34] and Massey,[36-38] we exploit nonstandard stochastic orderings on S [consistent with (28)] and stochastic monotonicity of the Markov process in this ordering in order to cope with the dimension of the state space. In particular, in our theorem, we shall assume that the transition function of the CTMC $Y_1(t)$ is stochastically monotone.[32-38]

Definition 1: A CTMC $Y_1(t)$ has a stochastically monotone transition function (kernel) $K_t \equiv K_t(s, A) \equiv P(Y_1(t) \in A \mid Y_1(0) = s)$ if $P_1K_t \leq P_2K_t$ in $(\mathscr{P}, \leq)$ whenever $P_1 \leq P_2$ in $(\mathscr{P}, \leq)$, where $(P_iK_t)(A) = \sum_{s \in S} P_i(s)K_t(s, A)$.

Remark 1: It is significant in Definition 1 that both the condition and the conclusion involve the same (unspecified) order relation $\leq$ on $\mathscr{P}$. □

Remark 2: As in Section 2 of Keilson and Kester,[33] stochastic monotonicity of a CTMC $Y_1(t)$ can be characterized by the transition rate function $q_1(s; A)$ and, after uniformization, by the transition function $I + \epsilon q_1$ of an associated discrete-time Markov chain with the same stationary distribution, where I is the identity map and ϵ is sufficiently small so that $I + \epsilon q_1$ is nonnegative. In particular, (i) $(P_1q_1)(A) \leq (P_2q_1)(A)$ for all $A \in \mathscr{A}$ whenever $P_1 \leq P_2$ and (ii) $P_1(I + \epsilon q_1) \leq P_2(I + \epsilon q_1)$ whenever $P_1 \leq P_2$ are each necessary and sufficient for $Y_1(t)$ to have a stochastically monotone transition function. □

For $A \subseteq S$, let $A^c = S - A$. Let $\hat{\pi}_2$ be the marginal distribution of π_2 on S, that is, $\hat{\pi}_2(A) = \pi_2(A \times S')$.

Theorem 6: Suppose that the CTMCs $Y_1(t)$ and $(Y_2(t), \Gamma(t))$ defined above have unique stationary distributions π_1 on S and π_2 on $S \times S'$. If (i) $Y_1(t)$ has a stochastically monotone transition function in $(\mathscr{P}, \leq)$ and (ii) for all $A \in \mathscr{A}$ and $\gamma \in S'$, $q_2(s, \gamma; A \times S') \leq q_1(s; A)$ for all $s \in A^c$ and $q_2(s, \gamma; A^c \times S') \geq q_1(s; A^c)$ for all $s \in A$, then $\hat{\pi}_2 \leq \pi_1$ in $(\mathscr{P}, \leq)$.

Proof: Since π_2 is the unique stationary distribution of $[Y_2(t), \Gamma(t)]$,

$$0 = (\pi_2 q_2)(A) = \sum_{s,\gamma} \pi_2(s, \gamma) q_2(s, \gamma; A)$$

for all $A \subseteq S \times S'$. By condition (ii),

$$0 = (\pi_2 q_2)(A \times S') \leq \sum_{s,\gamma} \pi_2(s, \gamma) q_1(s, A) = (\hat{\pi}_2 q_1)(A) \qquad (35)$$

for all $A \in \mathscr{A}$. Since the transition function associated with q_1 is stochastically monotone, (35) implies that $\hat{\pi} \leq \pi_1$. To see this, let P_{d1}

be the stochastically monotone transition function $I + \epsilon q_1$ of the associated discrete-time Markov chain constructed by uniformization. Then $0 \leq (\hat{\pi}_2 q_1)(A)$ for all $A \in \mathscr{A}$ is equivalent to $\hat{\pi}_2 \leq \hat{\pi}_2 P_{d1}$. Since P_{d1} is stochastically monotone, $\hat{\pi}_2 \leq \hat{\pi}_2 P_{d1} \leq \hat{\pi}_2 P_{d1}^2 \leq \cdots \leq \hat{\pi}_2 P_{d1}^n$. Since $\hat{\pi}_2 P_{d1}^n \to \pi_1$ as $n \to \infty$ and $\leq$ is a closed order, $\hat{\pi}_2 \leq \hat{\pi}_2 P_{d1}^n \leq \pi_1$. □

Remark 1: If $Y_2(t)$ is a Markov processes, so that we do not need $\Gamma(t)$, then Theorem 6 follows from Section 4.2 of Stoyan.[35] In fact, as explained in Ref. 31, Theorem 6 can also be viewed as a consequence of both Stoyan[35] and Massey.[38] □

Remark 2: For both Markov and non-Markov processes, the conditions of Theorem 6 also imply stochastic comparisons for the marginal distributions at time t for all t.[31] □

Remark 3: To relate Theorem 6 here to Theorem 5 of Ref. 30, note that it suffices to let one of the processes there, say $Y_1(t)$, have transition rates that do not depend on the extra information; that is, let $\lambda_1(k, I_t) = \alpha_1(k)$ and $\mu_1(k, I_t) = \beta_1(k)$. (The more general case follows by just making two comparisons.) Then $Y_1(t)$ becomes a birth-and-death process on the integers, which is known to be stochastically monotone with the usual stochastic order for probability measures on the real line. Theorem 6 here thus yields stochastic order (which is weaker than the MLR ordering in Ref. 30) under the conditions of the corollary to Theorem 5 in Ref. 30. Since the stationary distribution of $Y_1(t)$ depends on $\alpha_1(k)$ and $\beta_1(k+1)$ only through the rations $\alpha_1(k)/\beta_1(k+1)$, we can generalize the conditions here to the conditions of Theorem 5 in Ref. 30. In conclusion, then, Theorem 6 here yields a weaker conclusion (stochastic order instead of MLR order) under the same conditions as Theorem 5 of Ref. 30, but Theorem 6 here extends conveniently to the multivariate setting.

We now apply Theorem 6 to our problem. Theorem 1 follows immediately from (28), which we now establish.

Theorem 7: For each n-tuple $\mathbf{k} = (k_1, \cdots, k_n)$, $P(Y_i \leq k_i, 1 \leq i \leq n) \geq \prod_{i=1}^n P[N(s_i, \hat{\alpha}_i) \leq k_i]$.

Proof: We apply Theorem 6. The left and right sides of the inequality will be the stationary distributions of the processes $Y_2(t)$ and $Y_1(t)$, respectively, representing the number of busy servers at each facility for $1 \leq i \leq n$. In both cases, we assume that the service-time distributions are exponential, which we can do without loss of generality by Theorem 4. The process $Y_2(t)$ represents the process of interest to us and the process $Y_1(t)$ is a CTMC in which the coordinate stochastic processes are independent. In other words, $Y_1(t)$ is the process corresponding to n independent M/M/s/loss facilities. The information $\Gamma(t)$ associated with the process $Y_2(t)$ in Theorem 6 here is the number

of class j customers in service for each j at time t. It is easy to see that the process $\Gamma(t)$ and the bivariate process $[Y_2(t), \Gamma(t)]$ are CTMCs.

To fill in the rest of the details, let the state space S be the product of n integer intervals and let the state space S' for $\Gamma(t)$ be the product of c integer intervals, that is,

$$S = \overset{n}{\underset{i=1}{X}} \{0, 1, \cdots, s_i\} \quad \text{and} \quad S' = \overset{c}{\underset{i=1}{X}} \{0, 1, \cdots, \bar{s}\}, \tag{36}$$

where $\bar{s} = \max\{s_i, 1 \le i \le n\}$. Let S be endowed with the usual partial order in R^n; that is, $\mathbf{k}_1 \le \mathbf{k}_2$ for $\mathbf{k}_i = (k_{i1}, \cdots, k_{in})$ if $k_{1j} \le k_{2j}$ for all j. We shall be interested in lower subsets of S defined by

$$L(\mathbf{k}) = \{\mathbf{k}' \in S : \mathbf{k}' \le \mathbf{k}\}. \tag{37}$$

Let $\mathscr{A}$ be the set of complements of lower sets $L(\mathbf{k})$ for $\mathbf{k} \in S$; that is, $\mathscr{A} = \{L(\mathbf{k})^c \equiv S - L(\mathbf{k}) : \mathbf{k} \in S\}$. The set $\mathscr{A}$ induces a partial-order relation $\le$ on the space $\mathscr{P} \equiv \mathscr{P}(S)$ of all probability measures on S through the definition

$$P_1 \le P_2 \quad \text{if} \quad P_1(A) \le P_2(A) \quad \text{for all} \quad A \in \mathscr{A}. \tag{38}$$

Here $\le$ is a proper partial-order relation because $\mathscr{A}$ is a determining class.

It remains to show that conditions (i) and (ii) in Theorem 6 hold with respect to the ordering $\le$ in $\mathscr{P}(S)$. To see that condition (i) holds, that is, that q_1 is stochastically monotone with respect to $\le$, construct the associated discrete-time transition function $P_{d1} = I + \epsilon q_1$ (see Remark 2 before Theorem 6) and note that

$$(\pi P_{d1})[L(\mathbf{k})] = \sum_i p_i^{\pm} \pi[L(\mathbf{k} \pm \mathbf{e}_i)] + \left(1 - \sum_i p_i^{\pm}\right) \pi[L(\mathbf{k})], \tag{39}$$

where $\mathbf{e}_i$ is an n-tuple of all 0's except a 1 in one place and $p_i^{\pm}$ is a probability. (The permissible values of $\pm\mathbf{e}_i$ obviously depend on $\mathbf{k}$, but it is not necessary to specify them or the probabilities $p_i^{\pm}$ in detail.) From (39), it is immediate that $(\pi_1 P_{d1})[L(\mathbf{k})] \ge (\pi_2 P_{d1})[L(\mathbf{k})]$ for all $\mathbf{k} \in S$ if $\pi_1[L(\mathbf{k})] \ge \pi_2[L(\mathbf{k})]$ for all $\mathbf{k} \in S$.

To establish condition (ii), involving the comparison of the intensities, first apply Corollary 4.2 to make all the individual service rates identical without changing the stationary distributions being compared, as in the proof of Theorem 5. Next consider transitions upwards due to arrivals. Observe that for $\mathbf{k} \in L(\mathbf{k}')$

$$q_1[\mathbf{k}; L(\mathbf{k}')^c] = q_2[\mathbf{k}, \gamma; L(\mathbf{k}')^c] = 0 \tag{40}$$

unless $k_i = k_i'$ for some i and

$$q_2[\mathbf{k}, \gamma; L(\mathbf{k}')^c] \le q_1[\mathbf{k}; L(\mathbf{k}')^c] \tag{41}$$

otherwise. Make the comparison (41) by matching the intensities associated with each class j separately. For q_2 this corresponds to a simultaneous jump up of one in all coordinates of A_j with intensity λ_j, while for q_1 this corresponds to a jump up of one in one of the coordinates of A_j, each with intensity λ_j. Strict inequality occurs in (41) if the simultaneous transitions are blocked by the upper boundary, while the corresponding individual transition is not. Inequality also occurs if $k_i = k_i'$ for two or more indices i. Assuming that $k_i = k_i'$ for some i and there is no blocking at the upper boundary, the intensity of transition out of $L(\mathbf{k}')$ is λ_j for q_2 but $m\lambda_j$ for q_1, where m is the number of indices for which $k_i = k_i'$.

Next consider transitions downwards due to departures, where now all individual service rates are identical, say μ. (Invoke Corollary 4.2.) The transition function q_2 differs from q_1 by having multiple departures at intensity μ (that depend on the classes present) instead of individual departures each at intensity μ. The overall intensity of a transition downward, therefore, can be much greater in q_1, but with q_1 it is possible to enter the sets $L(\mathbf{k}')$ from outside, that is, from $\mathbf{k} \in L(\mathbf{k}')^c$ only by a departure in at most one of the coordinates. In other words, we have

$$q_2[\mathbf{k}, \gamma; L(\mathbf{k}')] \geq q_1[\mathbf{k}; L(\mathbf{k}')] \tag{42}$$

for all $\mathbf{k} \in L(\mathbf{k}')^c$. Strict inequality can occur in (42) if $k_i = k_i' + 1$ for two or more i in A_j and $k_i \leq k_i'$ otherwise when a class j customer is in service at time t. Then

$$q_2[\mathbf{k}, \gamma; L(\mathbf{k}')] = \mu > 0 = q_1[\mathbf{k}; L(\mathbf{k}')]. \tag{43}$$

for $\mathbf{k} \in L(\mathbf{k}')^c$. Properties (40) through (43) establish condition (ii) of Theorem 6 in our case. □

III. LARGE SYMMETRIC MODELS

To support the reduced-load approximation in Sections 1.5 and 1.6, we investigate large symmetric models. The limit theorems here are similar in spirit to previous ones for closed networks of queues with unlimited waiting space in Sections V and VIII in Ref. 40.

Here we assume that all facilities have s servers, all service-time distributions are exponential, all service rates are 1, all customer class arrival rates are λ, and all customers require service from m facilities. We associate one class with each possible subset of size m. We let the number of facilities n become large with the total offered load per facility $\hat{\alpha}$ held fixed. We achieve this by letting the arrival rate per class when there are n facilities be

$$\lambda_n = \hat{\alpha} n \Big/ \binom{n}{m} = \frac{\hat{\alpha} m!(n-m)!}{(n-1)!}. \tag{44}$$

It seems useful to focus on the stochastic process $Q_{nj}(t)$ representing the number of facilities with j busy servers at time t in the model with n facilities. Obviously, $Q_{n0}(t) = n - [Q_{n1}(t) + \cdots + Q_{ns}(t)]$ so that it suffices to focus on j with $1 \le j \le s$. The process $[Q_{n1}(t), \cdots, Q_{ns}(t)]$ is convenient because its dimension does not change as $n \to \infty$. It also appears that this process contains the essential information to characterize the asymptotic behavior of the blocking probability. However, this process presents a serious difficulty because, except in the relatively elementary special case in which $s = 1$, this process is not Markov. The future evolution of the process given any present value depends on additional information, namely, the specific classes present. However, we show that in a sense this information is asymptotically irrelevant.

3.1 A conjectured diffusion process limit

Let $V_{nj}(t)$ be the normalized stochastic process defined by

$$V_{nj}(t) = (Q_{nj}(t) - n\beta_j)/\sqrt{n}, \qquad t \ge 0, \tag{45}$$

and let $\mathbf{V}_n \equiv \mathbf{V}_n(t)$ be the vector-valued process defined by

$$\mathbf{V}_n(t) = [V_{n1}(t), \cdots, V_{ns}(t)], \qquad t \ge 0. \tag{46}$$

In the spirit of many limit theorems for closely related Markov processes,[61-63] we conjecture that $\mathbf{V}_n$ converges in distribution to a multivariate diffusion process. It should be possible to establish weak convergence (convergence in distribution) in the function space $D[0, \infty)$ of right-continuous functions with left limits,[61,64,65] but we support the diffusion approximation only by establishing convergence of the infinitesimal means. For the following conjecture, let $\mathbf{V}_n(t)$ be the stationary version (starting in equilibrium at $t = 0$) for each n, which exists and is unique by Theorem 4. The conjectured limit process is an s-dimensional multivariate Ornstein-Uhlenbeck diffusion process, which is characterized by its infinitesimal means and covariances.[62,63,66] The infinitesimal means and covariances have the relatively simple form of $M\mathbf{v}$ and Σ, where $\mathbf{v}$ is the s-dimensional state vector and M and Σ are $s \times s$ matrices that do not depend on the state.

Conjecture 5: The sequence of stationary stochastic process $\{\mathbf{V}_n, n \ge 1\}$ *defined in* (45) *and* (46) *converges weakly (in distribution) in the function space* $D([0, \infty), R^s)$ *to a stationary multivariate Ornstein-Uhlenbeck diffusion process if the normalization constants* β_j *in* (45) *are defined by* (10) *and* (11).

Heuristic Argument: In support of Conjecture 5, we prove that the infinitesimal means of $\{\mathbf{V}_n\}$ converge as $n \to \infty$ to those of an s-dimensional Ornstein-Uhlenbeck diffusion process. Even though the

process $\mathbf{V}_n(t)$ is not Markov for each n, the infinitesimal means depend on the past $\{\mathbf{V}_n(s), s \le t\}$ only through the present state $\mathbf{V}_n(t) = \mathbf{v}$ for each n. For $1 \le j \le s-1$, the infinitesimal means are

$$\begin{aligned}
&\mathbf{m}_{nj}(v_1, \cdots, v_s) \\
&\equiv \lim_{s \to 0} E\left[\frac{V_{nj}(t+s) - V_{nj}(t)}{s} \middle| \mathbf{V}_n(u),\ u \le t,\ \mathbf{V}_n(t) = \mathbf{v} \equiv (v_1, \cdots, v_s)\right] \\
&\approx n^{-1/2}\left\{(n\beta_{j-1} + \sqrt{n}v_{j-1})(m\alpha)\left(\frac{n - n\beta_s - \sqrt{n}v_s}{n}\right)^{m-1}\right. \\
&\quad + m(j+1)(n\beta_{j+1} + \sqrt{n}v_{j+1}) - (n\beta_j + \sqrt{n}v_j)(m\alpha) \\
&\quad \left.\cdot\left(\frac{n - n\beta_s - \sqrt{n}v_s}{n}\right)^{m-1} - mj(n\beta_j + \sqrt{n}v_j)\right\} \\
&\approx n^{1/2}\{\beta_{j-1}m\alpha(1-\beta_s)^{m-1} + m(j+1)\beta_{j+1} \\
&\quad - \beta_j m\alpha(1-\beta_s)^{m-1} - mj\beta_j\} + \{v_{j-1}m\alpha(1-\beta^s)^{m-1} \\
&\quad + v_{j+1}m(j+1) - v_j m\alpha(1-\beta_s)^{m-1} - v_j mj\}, \qquad (47)
\end{aligned}$$

where $\beta_0 = 1 - (\beta_1 + \cdots + \beta_s)$. For $j = s$, the infinitesimal mean is

$$\begin{aligned}
&\mathbf{m}_{ns}(v_1, \cdots, v_s) \\
&\approx n^{-1/2}\left\{(n\beta_{s-1} + \sqrt{n}v_{s-1})(m\alpha)\left(\frac{n - n\beta_s - \sqrt{n}v_s}{n}\right)^{m-1}\right. \\
&\quad \left. - ms(n\beta_s + \sqrt{n}v_s)\right\} \\
&\approx n^{1/2}\{\beta_{s-1}m\alpha(1-\beta_s)^{m-1} - ms\beta_s\} + \{v_{s-1}m\alpha(1-\beta_s)^{m-1} - v_s ms\}. \qquad (48)
\end{aligned}$$

In order for $\mathbf{m}_{nj}(v_1, \cdots, v_s)$ to converge as $n \to \infty$, it is necessary and sufficient to have the coefficients of $n^{1/2}$ vanish in the first terms of (47) and (48); that is, we need

$$\begin{aligned}
&\beta_{j-1}\alpha(1-\beta_s)^{m-1} + (j+1)\beta_{j+1} = \beta_j\alpha(1-\beta_s)^{m-1} + j\beta_j,\ j \le s-1, \\
&\qquad \beta_{s-1}\alpha(1-\beta_s)^{m-1} = s\beta_s. \qquad (49)
\end{aligned}$$

By induction, it follows that (10) and (11) provide the unique solution to (49). The remaining terms in (47) and (48) provide the infinitesimal means of the limiting diffusion process.

A next step to establish Conjecture 5 would be to establish convergence of the infinitesimal covariances, but the infinitesimal covariances do depend on more than the current state $\mathbf{v}$ for each n, and seem difficult to calculate. Finally, this would not actually complete the proof because the process $\mathbf{V}_n(t)$ is not Markov. [It almost would if $\mathbf{V}_n(t)$ were Markov by page 268 of Stroock and Varadhan.[63]]

Conjecture 6 (Corollary to Conjecture 5): The stationary random vector of $\mathbf{V}_n(t)$ is asymptotically normally distributed with zero mean vector as $n \to \infty$.

3.2 A law of large numbers

To establish Theorem 3 in Section 1.4, we prove a weaker result than Conjecture 5, namely, a functional law of large numbers for the process $\{[Q_{n1}(t), \cdots, Q_{ns}(t)], t \geq 0\}$ as $n \to \infty$. For this purpose, let

$$X_{nj}(t) = n^{-1}Q_{nj}(t), \qquad 1 \leq j \leq s, \tag{50}$$

and

$$\mathbf{X}_n(t) = [X_{n1}(t), \cdots, X_{ns}(t)] \tag{51}$$

for $t \geq 0$. Note that the components of $\mathbf{X}_n(t)$ are always nonnegative and their sum is at most one, so we can let the state space for $\mathbf{X}_n(t)$ be the s-dimensional simplex, say Δ, which is a compact subset of R^s.

The limiting stochastic process $\mathbf{X}(t)$ for $\mathbf{X}_n(t)$ will be a continuous deterministic motion, that is, a Markov diffusion process with zero diffusion or variance coefficient. The process $\{\mathbf{X}(t), t \geq 0\}$ has a transition function

$$P[\mathbf{X}(t) = T(t, \mathbf{x}) \mid \mathbf{X}(0) = \mathbf{x}] = 1,$$

where $\mathbf{x} \in \Delta$ and $T(t, \cdot)$ is a deterministic function mapping Δ into itself. Let $T_j(t, \mathbf{x})$ be the jth component of $T(t, \mathbf{x})$, that is, $T(t, \mathbf{x}) = [T_1(t, \mathbf{x}), \cdots, T_s(t, \mathbf{x})]$. The function $T(t, \cdot)$ is characterized by its derivative with respect to t, say $T'(\mathbf{x}) = [T_1'(\mathbf{x}), \cdots, T_s'(\mathbf{x})]$, where $T_j'(\mathbf{x}) = d/(dt)T_j(t, \mathbf{x})$, which is independent of t and is essentially the infinitesimal generator. Let $\Rightarrow$ denote weak convergence (convergence in distribution) of random elements in any space, for example, the state space Δ or the space of all sample paths $D([0, \infty), \Delta)$.[61,64,65]

Theorem 8: Assume exponentially distributed service times with mean one. If $\mathbf{X}_n(0) \Rightarrow \mathbf{X}(0)$ in Δ, then $\mathbf{X}_n \Rightarrow \mathbf{X}$ in $D([0, \infty), \Delta)$, where $\mathbf{X}(t)$ is a continuous deterministic motion with transition function $T(t, \mathbf{x})$ having derivatives with respect to t

$$T_j'(\mathbf{x}) = m[\hat{\alpha}(1 - x_s)^{m-1}(x_{j-1} - x_j) + (j + 1)x_{j+1} - jx_j], \; j \leq s - 1,$$

$$T_s'(\mathbf{x}) = m[\hat{\alpha}(1 - x_s)^{m-1}x_{s-1} - sx_s], \tag{52}$$

where $\mathbf{x} = (x_1, \cdots, x_s)$ and $x_0 = 1 - (x_1 + \cdots + x_s)$.

Proof: There are two steps, which we establish in Lemmas 1 and 2 below. First, we show that $\{\mathbf{X}_n\}$ is uniformly tight in $D([0, \infty), \Delta)$, so that every subsequence has a weakly convergent subsequence (see page

35 of Ref. 61). In the process, we show that each limit process has continuous sample paths. Then we show that the transition functions $P[\mathbf{X}_n(t_1 + t_2) \in A \mid X_n(t_1) = \mathbf{x}]$ converge to the transition function of the specified continuous deterministic motion as $n \to \infty$. Moreover, we show that the transition probability is asymptotically Markov, that is, asymptotically independent of the history of the process before t_1.

By Lemma 1, there is a weakly convergent subsequence, and any weakly convergent subsequence, say $\{\mathbf{X}_{n_k}\}$, has some limit process $\mathbf{X}'$. As a consequence of the weak convergence in the function space and the continuous mapping theorem (Theorem 5.1 of Billingsley[61]), the bivariate joint distributions converge weakly in Δ^2 too; that is,

$$P\{[\mathbf{X}_{n_k}(t_1), \mathbf{X}_{n_k}(t_2)] \in \cdot\} \Rightarrow P\{[\mathbf{X}'(t_1), \mathbf{X}'(t_2)] \in \cdot\}$$

for all $t_1, t_2 \geq 0$. Since $\mathbf{X}_n(0) \Rightarrow \mathbf{X}(0)$, $\mathbf{X}'(0)$ must be distributed the same as $\mathbf{X}(0)$. Moreover, since the transition functions converge, the limit $\mathbf{X}'(t)$ must be distributed as $T[t, \mathbf{X}(0)]$. Since the sample paths of $\mathbf{X}'$ are continuous, this determines the distribution of $\mathbf{X}'$ in $D([0, \infty), \Delta)$. Since the distribution of the limit of every weakly convergent subsequence of $\{\mathbf{X}_n\}$ in $D([0, \infty), \Delta)$ is determined, the entire sequence thus converges weakly to the determined limit, by Theorem 2.3 of Billingsley.[61] □

Lemma 1: The sequence $\{\mathbf{X}_n\}$ *is uniformly tight in* $D([0, \infty), \Delta)$ *and the limit of any convergent subsequence has continuous paths.*

Proof: To establish uniform tightness in $D([0, \infty), \Delta)$, we establish the stronger C-tightness, conditions for which are given in Theorem 8.3 of Billingsley.[61] This implies that $\{\mathbf{X}_n\}$ is also D-tight and that the limit of any convergent subsequence has continuous sample paths. To establish C-tightness, it suffices to focus on a single coordinate of $\{\mathbf{X}_n\}$ in $D([0, \infty), R)$, say $\{X_{nj}\}$ (see Section 2 of Ref. 65 and Exercise 6, page 41, of Ref. 61). Moreover, it suffices to restrict the time interval to a compact subinterval.[64,65,67] Since the state space Δ of $\mathbf{X}_n$ is a compact subset of R^s, the set of all probability measures on Δ with the topology of weak convergence is metrizable as a compact metric space (see page 45 of Ref. 68). By Prohorov's theorem, page 37 of Ref. 61, $\{X_{nj}(0)\}$ is uniformly tight in R and condition (i) of Theorem 8.3 in Billingsley[61] holds.

We establish the remaining condition (ii) by bounding the change in $X_{nj}(t)$ in a fixed interval of length δ by the normalized sum of all arrivals and all departures during that arrival. The arrivals, in turn, are bounded by the total number of arrivals that would occur if all servers remained empty throughout the interval, that is, by a Poisson random variable with rate $nm\hat{\alpha}\delta$. Similarly, the number of departures is bounded above by the number of departures that would occur if all facilities remained full throughout the interval, that is, by a Poisson

random variable with rate $nms\delta$. These two bounds can be expressed via stochastic order relations, as in (23), by actually generating the arrivals and departures by appropriately thinning two independent Poisson processes with the indicated rates.[69]

To establish condition (*ii*), it remains to show that for all positive c, ϵ, and η there exists δ such that

$$P[N(cn\delta) > n\epsilon] < \delta\eta \tag{53}$$

for all n sufficiently large, where $N(\lambda)$ is a Poisson random variable with mean λ. Of course, we choose δ so that $c\delta < \epsilon$ to have the mean of $N(cn\delta)$ less than $\eta\epsilon$. Then, using Chebyschev's inequality, we obtain

$$\begin{aligned} P[N(cn\delta) > n\epsilon] &< \frac{\text{Var } N(cn\delta)}{[EN(cn\delta) - n\epsilon]^2} \\ &< \frac{cn\delta}{(cn\delta - n\epsilon)^2} = \frac{c\delta}{n(c\delta - \epsilon)^2}, \end{aligned}$$

which shows that (53) indeed holds for all $n > n_0$, where $n_0 = c/[\eta(c\delta - \epsilon)]$. □

Let A^ϵ be the open ϵ-ball in Δ about the set A, that is,

$$A^\epsilon = \{\mathbf{x} \in \Delta: d(\mathbf{x}, \mathbf{y}) < \epsilon \quad \text{for some} \quad \mathbf{y} \in A\}, \tag{54}$$

where d is a metric on R^s, here taken to be the maximum metric $d(\mathbf{x}, \mathbf{y}) = \max\{|x_i - y_i|, 1 \le i \le s\}$.

Lemma 2: For all positive ϵ, states $\mathbf{x} \in \Delta$ and histories $\{\mathbf{X}_n(u), u \le t_1\}$,

$$\lim_{n\to\infty} P(\mathbf{X}_n(t_1 + t_2) \in [T(t_2, \mathbf{x})]^\epsilon \,|\, \mathbf{X}_n(u), u \le t_1, \mathbf{X}_n(t_1) = \mathbf{x}) = 1,$$

where T is the continuous deterministic motion in Theorem 8.

Proof: Let I be the identity map on Δ. Since $T(t, \cdot)$ has the semigroup property of a Markov process and the derivative T' is bounded and continuous, $(I + \epsilon T')^{t/\epsilon} \to T(t, \cdot)$ as $\epsilon \to 0$. Consequently, it suffices to prove that there is a constant K such that for all sufficiently small positive ϵ

$$\lim_{n\to\infty} P(\mathbf{X}_n(t + \epsilon) \in (\mathbf{x} + \epsilon T')^{K\epsilon^2} \,|\, \mathbf{X}_n(u), u \le t, \mathbf{X}_n(t) = \mathbf{x}) = 1. \tag{55}$$

To establish (55), we use stochastic dominance arguments as in the proof of Lemma 1. In particular, we first observe that, for any n, t and ϵ, the total number of arrivals in the interval $[t, t + \epsilon]$ is stochastically dominated by a Poisson variable with mean $nm\hat{a}\epsilon$. Similarly, for any n, t, and ϵ, the total number of departures in the interval $[t, t + \epsilon]$ is stochastically dominated by a Poisson variable with mean $nms\epsilon$. These stochastic bounds give us initial bounds on how much $\mathbf{X}_n(u)$ can differ from $\mathbf{X}_n(t)$ in the interval $[t, t + \epsilon]$ for all possible histories. Since

$X_{nj}(t)$ is a proportion, we can apply a law of large numbers for Poisson variables as the rate increases. In particular, there is a constant K, which is independent of ϵ as $\epsilon \to 0$, such that

$$\lim_{n\to\infty} P\Big(\sup_{t\le u\le t+\epsilon} |X_{nj}(u) - X_{nj}(t)| > K\epsilon \,|\, \mathbf{X}_n(u), \quad u \le t, \quad \mathbf{X}_n(t) = \mathbf{x} \Big) = 0. \quad (56)$$

We now use the initial bound in (56) to produce better bounds on $\mathbf{X}_n(t+\epsilon) - \mathbf{X}_n(t)$, that is, to establish (55). Given that $\mathbf{X}_n(t) = \mathbf{x}$ and

$$\sup_{t\le u\le t+\epsilon} |X_{nj}(u) - X_{nj}(t)| < K\epsilon$$

for $1 \le j \le s$, the actual flow rate into state j (the rate of increase of $X_{nj}(u)$) in the interval $[t, t+\epsilon]$ is bounded above by

$$\begin{aligned} I^u(j) &= \hat{\alpha} \min\{1, (1 - x_s + K\epsilon)^{m-1}\}(x_{j-1} + K\epsilon) \\ &\quad + (j+1)(x_{j+1} + K\epsilon) \\ &\le \hat{\alpha} \min\{1, (1 - x_s + K\epsilon)^{m-1}\}x_{j-1} + \hat{\alpha}K\epsilon + (j+1)x_{j+1} \\ &\quad + (j+1)K\epsilon \\ &\le \hat{\alpha}(1 - x_s)^{m-1}x_{j-1} + (j+1)x_{j+1} + (\hat{\alpha}m + (j+1))K\epsilon \quad (57) \end{aligned}$$

and bounded below by

$$\begin{aligned} I^l(j) &= \hat{\alpha} \max\{0, (1 - x_s - K\epsilon)^{m-1}\}(x_{j-1} - K\epsilon) \\ &\quad + (j+1)(x_{j+1} - K\epsilon) \\ &\ge \hat{\alpha}(1 - x_s)^{m-1}x_{j-1} + (j+1)x_{j+1} - [\hat{\alpha}m + (j+1)]K\epsilon. \quad (58) \end{aligned}$$

In other words, with n facilities the flow into state j for the unnormalized process $Q_{nj}(t)$ is stochastically bounded above by a Poisson process with rate $nI^u(j)$ and stochastically bounded below by a Poisson process with rate $nI^l(j)$.[69] Similarly, the flow rate out of state j [the rate of decrease of $X_{nj}(u)$] in the interval $[t, t+\epsilon]$ is bounded above by

$$\begin{aligned} O^u(j) &= \hat{\alpha} \min\{1, (1 - x_s + K\epsilon)^{m-1}\}(x_j + K\epsilon) + j(x_j + K\epsilon) \\ &\le \hat{\alpha}(1 - x_s)^{m-1}x_j + jx_j + (\hat{\alpha}m + j)K\epsilon \quad (59) \end{aligned}$$

and bounded below by

$$\begin{aligned} O^l(j) &= \hat{\alpha} \max\{0, (1 - x_s - K\epsilon)^{m-1}\}(x_j - K\epsilon) + j(x_j - K\epsilon) \\ &\ge \hat{\alpha}(1 - x_s)^{m-1}x_j + jx_j - (\hat{\alpha}m + j)K\epsilon. \quad (60) \end{aligned}$$

We invoke a well-known functional law of large numbers for the

Poisson process (which is a consequence of the functional central limit theorem, Section 17 of Billingsley[61]) to deduce that as $n \to \infty$ the change in $X_n(u)$, that is, the change in the proportions, is bounded above and below by the deterministic motions with rates $I^u(j) - O^l(j)$ and $I^l(j) - O^u(j)$, respectively. Hence, for any history $\{\mathbf{X}_n(u), u \le t\}$ and any state $\mathbf{X}_n(t) = \mathbf{x}$,

$$\lim_{n\to\infty} P\{\epsilon[I^l(j) - O^u(j)] \le X_{nj}(t+\epsilon) - X_{nj}(t) \le \epsilon[I^u(j) - O^l(j)] \,|\, \mathbf{X}_n(u), u \le t, \mathbf{X}_n(t) = \mathbf{x}\} = 1, \quad (61)$$

but

$$\epsilon[I^u(j) - O^l(j)] = \epsilon T_j'(\mathbf{x}) + \epsilon^2 K'$$

and

$$\epsilon[I^l(j) - O^u(j)] = \epsilon T_j'(\mathbf{x}) - \epsilon^2 K'$$

for $K' = (2\hat{\alpha}m + 2j + 1)K$, so that (61) is equivalent to the desired result. □

We now describe the limiting continuous deterministic motion $\mathbf{X}(t)$ specified in Theorem 8. In particular, we verify that $\mathbf{X}(t)$ has a unique stationary distribution and converges to it as $t \to \infty$ for any initial distribution. It is relatively elementary that $T(t, \cdot)$ has a unique fixed point in Δ. We want to establish the stronger result that $T(t, \cdot)$ has a unique fixed point in the space $\mathscr{P}(\Delta)$ of all probability measures on Δ. To appreciate the difference, note that clockwise circular motion at constant angular velocity in the plane has a unique fixed point in the plane, namely, the origin, but the uniform distribution over any circle centered about the origin is a stationary distribution for this clockwise circular motion. We show that our continuous deterministic motion actually converges to its unique fixed point in Δ for every initial distribution.

Theorem 9: For any initial vector $\mathbf{y}$, $\mathbf{X}(t) \to \beta$ *as* $t \to \infty$, *where* $\beta \equiv (\beta_1, \cdots, \beta_s)$ *is determined by* (10) *and* (11).

Corollary 9.1: The limiting continuous deterministic motion $\mathbf{X}(t)$ *has a unique stationary distribution, which is a unit mass on the vector* β *determined by* (10) *and* (11).

Proof: We write $T_t(A_1) \to A_2$ as $t \to \infty$ for subsets A_1 and A_2 of Δ to represent that $T(t, \mathbf{y}) \to A_2$ as $t \to \infty$ for all $\mathbf{y} \in A_1$, that is, $d(T(t, \mathbf{y}), A_2) \to 0$ as $t \to \infty$, where $d(\mathbf{x}, A) = \inf\{d(\mathbf{x}, \mathbf{y}): \mathbf{y} \in A\}$ with d the metric on R^{s-1}. Equivalently, $T(t, \mathbf{y}) \to A_2$ as $t \to \infty$ if the limits of all convergent subsequences $\{T(t_k, \mathbf{y}), k = 1, 2, \cdots\}$ of $\{T(t, \mathbf{y}), t \ge 0\}$ with $t_k \to \infty$ are contained in A_2. (Since Δ is a compact metric space, every sequence has a convergent subsequence. Moreover, the limit sets

A_2 considered below will be closed, so that they will contain the limits.) Our goal is to show that $T_t(\Delta) \to \{\beta\}$. To do so, we construct compact subsets $L_1, \cdots, L_s$ such that

$$L_s = \{\beta\} \subseteq L_{s-1} \subseteq \cdots \subseteq L_1 \subseteq \Delta, \tag{62}$$

$T_t(\Delta) \to L_1$ and $T_t(L_k) \to L_{k+1}$, $1 \leq k \leq s-1$, as $t \to \infty$. Since T_t has the semigroup property $T(t_1 + t_2, \mathbf{x}) = T[t_1, T(t_2, \mathbf{x})]$ for all $\mathbf{x}$, t_1 and t_2, and is continuous, this implies that $T_t(\Delta) \to L_k$ for all k, so that $T_t(\Delta) \to \{\beta\}$.

We consider real-valued functionals of $T(t, \cdot)$. First we consider the net flow into the set $\{1, \cdots, s\}$, defined by

$$F_{st}(\mathbf{x}) = \sum_{j=1}^{s} T_j(t, \mathbf{x})$$

with derivative

$$F_s'(\mathbf{x}) = \hat{\alpha}(1 - x_s)^m - \sum_{j=1}^{s} jx_j, \tag{63}$$

which is continuous and strictly decreasing in $\mathbf{x}$. Moreover, for all $\mathbf{x}$ sufficiently large, $F_s'(\mathbf{x}) < 0$; and for all $\mathbf{x}$ sufficiently small, $F_s'(\mathbf{x}) > 0$. Consequently, $F_{st}(\mathbf{x}) \to 0$, $F_s'[T(t, \mathbf{x})] \to 0$ and $T_t(\Delta) \to L_1$ as $t \to \infty$, where

$$L_1 = \{\mathbf{x} \in \Delta: F_s'(\mathbf{x}) = 0\}. \tag{64}$$

Next consider the net flow into the states $\{1, \cdots, s-1\}$, defined by

$$F_{(s-1),t}(\mathbf{x}) = \sum_{j=1}^{s-1} T_j(t, \mathbf{x})$$

with derivative

$$\begin{aligned} F_{s-1}'(\mathbf{x}) &= \hat{\alpha}(1 - x_s)^{m-1}(1 - x_s - x_{s-1}) - \sum_{j=1}^{s-1} jx_j \\ &= \left[\hat{\alpha}(1 - x_s)^m - \sum_{j=1}^{s} jx_j\right] + [sx_s - x_{s-1}\hat{\alpha}(1 - x_s)^{m-1}] \\ &= F_s'(\mathbf{x}) + [sx_s - x_{s-1}\hat{\alpha}(1 - x_s)^{m-1}]. \end{aligned} \tag{65}$$

For $\mathbf{x} \in L_1$, $F_s'(\mathbf{x}) = 0$, and $F_{s-1}'(\mathbf{x}) = [sx_x - x_{s-1}\hat{\alpha}(1 - x_s)^{m-1}]$, which is continuous and strictly decreasing in $(x_{s-1}, -x_s)$. For all $\mathbf{x} \in L_1$ with x_{s-1} sufficiently large (small) and x_s sufficiently small (large), $F_{s-1}'(\mathbf{x}) < 0$ (>0). Hence, $F_{(s-1)t}(\mathbf{x}) \to 0$ and $F_{s-1}'[T(t, \mathbf{x})] \to 0$ for $\mathbf{x} \in L_1$, and $T_t(L_1) \to L_2$ as $t \to \infty$, where

$$L_2 = \{\mathbf{x} \in L_1: F_{s-1}'(\mathbf{x}) = 0\}. \tag{66}$$

Similarly, we consider the net flow $F_{(s-2)t}(\mathbf{x})$ into the states $\{1, \cdots, s-2\}$ with derivative

$$F'_{s-2}(\mathbf{x}) = F'_s(\mathbf{x}) + F'_{s-1}(\mathbf{x}) + (s-1)x_{s-1} - x_{s-2}\hat{\alpha}(1-x_s)^{m-1}, \quad (67)$$

which is continuous and strictly decreasing in $(x_{s-2}, x_{s-1}, -x_s)$. Moreover, for all $\mathbf{x} \in L_2$ with x_{s-2} sufficiently large (small) and x_{s-1} and x_s sufficiently small (large), $F'_{s-2}(\mathbf{x}) < 0$ (>0). Hence, $T_t(L_2) \to L_3$, where

$$L_3 = \{\mathbf{x} \in L_2 : F'_{s-2}(\mathbf{x}) = 0\}. \quad (68)$$

The proof is completed by induction. The s equations $F'_k(\mathbf{x}) = 0$, $1 \le k \le s$, uniquely determine the fixed point β of $T(t, \cdot)$ in Δ defined by (10) and (11). These are the partial balance equations for a single M/M/s/loss facility.[6] Hence, $L_s = \{\beta\}$ and $T_t(\Delta) \to \{\beta\}$ as $t \to \infty$. □

3.3 Proof of Theorem 3(a)

Proof: We now apply Theorems 8 and 9 to prove Theorem 3(a). Let $\mathbf{Z}_n = (Z_{n1}, \cdots, Z_{ns})$ have the unique stationary distribution of $\{\mathbf{X}_n(t), t \ge 0\}$ for each n. (Existence and uniqueness follow from Theorem 4.) Since the state space for $\{\mathbf{Z}_n\}$ is the compact simplex Δ in R^s, the sequence $\{\mathbf{Z}_n\}$ is uniformly tight and has a weakly convergent subsequence, say $\{\mathbf{Z}_{n_k}\}$; apply the argument in the proof of Theorem 8. Since $\mathbf{Z}_{n_k} \Rightarrow \mathbf{Z}$ in Δ as $n_k \to \infty$ for some $\mathbf{Z}$, the stationary versions of the stochastic processes $\mathbf{X}_{n_k}(t)$ converge weakly (in distribution) in $D([0, \infty), \Delta)$ and $n_k \to \infty$ to the continuous deterministic motion $\mathbf{X}(t)$ with $\mathbf{X}(0)$ distributed as $\mathbf{Z}$ (applying Theorem 8). However, since $\mathbf{X}_{n_k}(t)$ is stationary for each n_k, so is $\mathbf{X}(t)$. By Corollary 9.1, the only stationary distribution for $\mathbf{X}(t)$ is the limiting vector β. Hence, we must have $P(\mathbf{Z} = \beta) = 1$. Since every convergent subsequence of $\{\mathbf{Z}_n\}$ has the same limit $\mathbf{Z}$, we must have convergence of the entire sequence, that is, $\mathbf{Z}_n \Rightarrow \mathbf{Z}$ in Δ (see Theorem 2.3 of Ref. 61). Since $P(\mathbf{Z} = \beta) = 1$ for the deterministic vector β, we have convergence in probability (see page 25 of Ref. 61). □

Remark: Theorems 3, 8, and 9 together imply that the stationary versions of the stochastic processes $\mathbf{X}_n(t)$ also satisfy a functional law of large numbers in $D([0, \infty), \Delta)$.

3.4 Proof of Theorem 3(b)

The key to Theorem 3(b), of course, is Theorem 3(a) and the symmetry: Every subset of size m is equally likely to be the set of m required facilities for each arrival. In addition to Theorem 3(b) we establish a stronger form of asymptotic independence, for the stochastic processes instead of only the stationary distributions. Let $Y_{ni}(t)$ be the number of busy servers at facility i at time t. Let $Y_n(t) = [Y_{n1}(t), \cdots, Y_{nn}(t)]$ be the stationary version for each n.

Theorem 10: For any finite subset H and any t_0, the stationary stochastic processes $\{Y_{ni}(t), 0 \le t \le t_0\}$, $i \in H$, are asymptotically independent as $n \to \infty$.

Proof: By symmetry, the joint distribution of $\{Y_{ni}(t), i \in H\}$ is invariant under a permutation of the indices. By Theorem 3a, the proportion of facilities with j busy servers converges in probability to β_j as $n \to \infty$. Hence, by symmetry, $\lim_{n\to\infty} P[Y_{ni}(0) = j_i, 1 \le i \le H] = \prod_{i\in I} \beta_{j_i}$, so that the initial stationary values $Y_{ni}(0)$, $i \in I$, are asymptotically mutually independent. Next, let $A_{ni}(t)$ be the arrival process to facility i excluding losses due to blocking elsewhere. By Theorem 3a, $A_{ni}(t)$ converges to a Poisson process with rate $\hat{\alpha}(1 - \beta_s)^{m-1}$ as $n \to \infty$. Moreover, again by symmetry and Theorem 3a, the arrival processes $\{A_{ni}(t), 0 \le t \le t_0\}$, $i \in H$, are asymptotically mutually independent as $n \to \infty$. Since probability that the facilities in H share any customers at any time in the interval $[0, t_0]$ is asymptotically negligible as $n \to \infty$, the departure processes for $i \in H$ and thus also the processes $\{Y_{ni}(t), 0 \le t \le t_0\}$, $i \in H$, are asymptotically mutually independent. □

IV. EXISTENCE, UNIQUENESS, AND INSENSITIVITY

We now prove Theorem 4.

Proof: In the case of exponentially distributed service times, the vector-valued stochastic process, say $[N_1(t), \cdots, N_c(t)]$, representing the number of class j customers in service at time t for all j, $1 \le j \le c$, is an irreducible c-dimensional continuous-time Markov chain with a finite state space. Hence, there exists a unique stationary distribution. It is easy to see that the claimed distribution in Theorem 4 is the steady-state distribution by making the standard partial balance analysis.[6-8] The same steady-state distribution holds for general service-time distributions by the insensitivity results, which we discuss further below.

To prove the rest of Theorem 4, we need to establish that the steady-state distribution of $(N_1, \cdots, N_c)$ is actually well defined. For this purpose, we construct a continuous-time vector-valued Markov process $\{\mathbf{Z}(t), t \ge 0\}$, depicting the number of class j customers in service for each j and the remaining service time of each at time t. [$\mathbf{Z}(t)$ is the continuous-time Markov process associated with the GSMP in Ref. 46.] The steady-state distribution in Theorem 4 is understood to be the marginal distribution corresponding to $(N_1, \cdots, N_c)$ of the stationary distribution of $\mathbf{Z}(t)$. We shall show that $\mathbf{Z}(t)$ indeed has a stationary distribution (without establishing uniqueness) and that the marginal distribution corresponding to $(N_1, \cdots, N_c)$ is always as claimed in Theorem 4 and so is unique.

For our given general service-time distributions, we construct se-

quences of approximating service-time distributions from finite mixtures of finite convolutions of exponential distributions, as in Section 3.3 of Ref. 6 and in the proof of Theorem 2 in Ref. 46. We construct this so that the means are unchanged and there is weak convergence to the given distributions. For our special model, it is easy to see that each continuous-time Markov process $\mathbf{Z}(t)$ so created with these approximating service-time distributions has a unique invariant probability measure. Existence follows from the theory of continuous-time Markov chains with finite-state space. Uniqueness follows from the irreducibility that is evident from our special structure. Moreover, the partial balance property satisfied by the steady-state distribution in the exponential case implies that the unique stationary distribution of $\mathbf{Z}(t)$ in each approximating case has marginal distribution for $(N_1, \cdots, N_c)$ as specified in Theorem 4.[8,44,45] Finally, we treat the case of the original general service-time distributions by continuity, invoking Theorem 3 of Ref. 46. (Note that uniqueness with the approximating service-time distributions is crucial for that theorem.) This continuity theorem implies that the process $\mathbf{Z}(t)$ indeed has a stationary distribution and that the marginal distribution corresponding to $(N_1, \cdots, N_c)$ is as claimed for every stationary distribution of $\mathbf{Z}(t)$. □

Remark: An alternate proof of existence and uniqueness can be constructed using the fact that arrival epochs when the system is empty constitute regeneration points. The GSMP theory is also useful for describing steady state in more general models for which this is not the case; for example, if the service-time distributions are nonexponential and the arrival processes for the different classes are independent non-Poisson renewal processes. However, the insensitivity is typically lost with this extension.

V. CONVERGENCE OF THE SUCCESSIVE APPROXIMATION ALGORITHM

Example 3 in Section 1.5 showed that the successive approximation scheme (8) need not converge. In this section we show that if the offered loads are sufficiently small, then the operator T defined by the right side of (7) is a contraction operator, so that it has a unique fixed point to which successive iterates of T converge geometrically fast. However, the conditions for this property are quite strong, so that the theorem does not nearly cover all practical cases.

To state our results, let $\|\cdot\|$ be the supremum norm on R^n defined by $\|\mathbf{x}\| = \max\{|x_i| : 1 \le i \le n\}$ for $\mathbf{x} = (x_1, \cdots, x_n)$. Let $\tilde{\alpha}_i(\mathbf{b})$ be the reduced offered load as a function of $\mathbf{b} \equiv (b_1, \cdots, b_n)$ as defined in (6). Let $\gamma(\mathbf{b})$ be defined by

$$\gamma(\mathbf{b}) = \max_{1 \le i \le n} \left\{ \left(\frac{s_i}{\tilde{\alpha}_i(\mathbf{b})} - 1 + b_i \right) b_i n \hat{\alpha}_i \right\}. \tag{69}$$

Let $\mathbf{U} \equiv (U_1, \cdots, U_n)$ be an upper bound on any solution $\mathbf{b}^*$ of (7) such as $(B(s_1, \hat{\alpha}_1), \cdots, B(s_n, \hat{\alpha}_n)) = T^2(1)$ or $T^{2k}(1)$ for any $k \geq 1$.

Theorem 11: If $\gamma(\mathbf{U}) < 1$ *for* γ *in* (69) *and the upper bound* $\mathbf{U}$ *to any solution of* (7), *then*

(*i*) $\| T(\mathbf{b}^1) - T(\mathbf{b}^2) \| \leq \gamma(\mathbf{U}) \| \mathbf{b}^1 - \mathbf{b}^2 \|$ *for all* $\mathbf{b}^1$ *and* $\mathbf{b}^2$ *in* R^n *with* $0 \leq b_i^1, b_i^2 \leq U_i$ *for all* i, *so that*

(*ii*) T *has a unique fixed point* $\mathbf{b}^*$ *in* $[\mathbf{0}, \mathbf{U}] \equiv \{\mathbf{b}: 0 \leq b_i \leq U_i\}$, *and*

(*iii*) $\| T^k(\mathbf{b}^0) - \mathbf{b}^* \| \leq \gamma(\mathbf{U})^k \| \mathbf{b}^0 - \mathbf{b}^* \|$ *for all* k *when the initial vector* $T^0(\mathbf{b}) = \mathbf{b}^0$ *is in* $[\mathbf{0}, \mathbf{U}]$.

Proof: Parts (ii) and (iii) follow from (i) by the Banach-Picard fixed-point theorem for a contraction map on a complete metric space.[70] For (i) it suffices to have

$$\left| \frac{\partial T_i(\mathbf{b})}{\partial b_k} \right| \leq \frac{\gamma(\mathbf{U})}{n}$$

for all i and k (for example, see Theorem 2, page 111 of Ref. 70). By Theorem 15 of Jagerman,[27]

$$\frac{\partial B(s, \alpha)}{\partial \alpha} = \left[\frac{s}{\alpha} - 1 + B(s, \alpha) \right] B(s, \alpha).$$

Hence,

$$\left| \frac{\partial T_i(\mathbf{b})}{\partial b_k} \right| \leq \left[\frac{s_i}{\bar{\alpha}_i(\mathbf{b})} - 1 + b_i \right] b_i \hat{\alpha}_i \leq \frac{\gamma_i(\mathbf{U})}{n}$$

for $\mathbf{b} \in [\mathbf{0}, \mathbf{U}]$.

Remark 1: For the symmetric model, (69) simplifies to

$$\gamma(U) = \frac{nsU}{(1 - U)^{m-1}} - (1 - U)Un\hat{\alpha}, \tag{70}$$

so that a simple sufficient condition for the condition of theorem 11 is

$$\frac{nsU}{(1 - U)^{m-1}} < 1. \tag{71}$$

Remark 2: If $U_i = B(s_i, \hat{\alpha}_i)$ for all i or if $\mathbf{U} = T^{2k}(1)$ for some k using (8), then U is an increasing function of the offered loads $(\hat{\alpha}_1, \cdots, \hat{\alpha}_n)$ or $(\alpha_1, \cdots, \alpha_c)$ because T is an increasing function of $\mathbf{b}$ and $B(s, \alpha)$ is an increasing function of α. Hence, if the offered loads are sufficiently small, then the vector $\mathbf{U}$ will be sufficiently small, so that the condition of Theorem 11 will eventually hold.

VI. CONCLUSIONS

We have investigated a model to describe the blocking probabilities

when service is required from several multiserver facilities simultaneously. We have shown in Theorem 1 that some standard approximations produce upper bounds. In the process, we have established several other useful stochastic comparison results (Theorems 5 through 7 and Ref. 31). We also have proposed an improved reduced-load approximation and developed an efficient algorithm (Theorem 2) to treat both the Poisson arrival case (Section 1.5) and the non-Poisson arrival case (Section 1.9). In Theorem 8 we have established a functional law of large numbers that implies that the symmetric reduced-load approximation is asymptotically correct for symmetric models as the number of facilities increases with the offered load per facility and the number of facilities per class held fixed (Theorems 3 and 10). We have displayed the exact formula in Theorem 4 and justified the insensitivity with respect to the service-time distributions (Sections 1.8 and IV).

Among the important directions for future research are (i) testing the approximations further, especially for non-Poisson arrival processes; (ii) establishing better conditions for the reduced-load equations (7) to have a unique solution (Conjectures 1 and 2); (iii) establishing better conditions for the successive approximation scheme (8) to converge; (iv) establishing lower bounds on the exact blocking probabilities paralleling the upper bounds in Theorem 1; (v) determining if the reduced-load approximation is an upper bound on the exact blocking probability for symmetric models (Conjecture 3); (vi) determining if the exact blocking probabilities for symmetric models are increasing in n when the offered load per facility is fixed (Conjecture 4); (vii) establishing (if possible) the diffusion limit in Section III (Conjectures 5 and 6); (viii) seriously analyzing smaller models in which the basic facility-independence approximation in (5) underlying all the approximations here is not appropriate.[9] In particular, in the spirit of Kaufman[7] and Mitra and Weinberger,[21] it would be nice to develop an efficient algorithm for the exact blocking probabilities in Theorem 4 and Corollary 4.1.

It would also be of interest to consider other related models, for example, models in which more than one server per facility may be required, and related delay systems. There are two kinds of waiting to be considered for delay systems: waiting for each customer class outside the system, and waiting for service at each facility within the system. The second form of waiting may still require simultaneous service or some other form.[14]

VII. ACKNOWLEDGMENTS

This work was initially motivated by discussions with D. D. Sheng about the PANDA software package.[2,3] I am grateful to her, D. P.

Heyman (Bell Communications Research), and D. R. Smith for initial stimulating discussions. The product bound in Theorem 1 was also conjectured by them. The summation bound in Corollary 1.2 was proved in the special case of two facilities by different methods by Sheng and Smith (see the appendix of Ref. 3). Symmetric models were apparently first investigated by Mitra and Weinberger,[21] but they were brought to my attention by Heyman and Smith. Heyman and Smith also developed the reduced-load approximation (9) for symmetric models and conjectured Theorem 3. I am grateful to W. A. Massey for showing me his unpublished work[38] and for commenting on Ref. 31, which presents the general stochastic comparison theory behind Theorems 1, 6, and 7. Finally, I am grateful to W. J. Hery and J. T. Wittbold (AT&T Communications) for discussions about their applications and the performance of the approximations.[24,25] They each programmed the reduced-load approximation (5) through (8) and investigated numerical examples. Tables IV and V come from Wittbold.

VIII. EPILOGUE

This section has been added in proof to report important new work. Kelly[71] has proved that the reduced-load system of eq. (7) has a unique solution, thus confirming Conjectures 1 and 2. Kelly also has proved that the reduced-load approximation is asymptotically correct in heavy traffic, that is, in a network with fixed topology in which $\alpha_j \to \infty$ and $s_i \to \infty$, as in Ref. 57. In fact, Kelly's heavy-traffic limit theorem is a multifacility generalization of the local limit theorem in the Appendix of Ref. 57.

Ziedens and Kelly[72] also have proved limit theorems similar to Theorem 3 for symmetric networks in which the number of nodes increases. For the special tree networks in Fig. 1, Mitra[73] has determined an efficient algorithm for the exact solution based on asymptotic expansions, in the spirit of Ref. 21. Other related work appears in Refs. 74 through 76.

REFERENCES

1. D. Bear, *Principles of Telecommunication-Traffic Engineering*, London: The Institution of Electrical Engineers, 1976.
2. D. D. Sheng, "Performance Analysis Methodology for Packet Network Design," IEEE Global Telecommun. Conf., *GLOBECOM '83* (December 1983), pp. 456–60.
3. C. L. Monma and D. D. Sheng, unpublished work.
4. E. Cinlar, *Introduction to Stochastic Processes*, Englewood Cliffs, New Jersey: Prentice-Hall, 1975.
5. R. W. Wolff, "Poisson Arrivals See Time Averages," Oper. Res., *30*, No. 2 (March–April 1982), pp. 223–31.
6. F. P. Kelly, *Reversibility in Stochastic Networks*, New York: Wiley, 1979.
7. J. S. Kaufman, "Blocking in a Shared Resource Environment," IEEE Trans. Commun., *COM-29*, No. 10 (October 1981), pp. 1474–81.
8. D. Y. Burman, J. P. Lehoczky, and Y. Lim, "Insensitivity of Blocking Probabilities

in a Circuit-Switching Network," J. Appl. Probab., *21,* No. 4 (December 1984), pp. 850–59.
9. J. M. Holtzman, "Analysis of Dependence Effects in Telephone Trunking Networks," B.S.T.J., *50,* No. 8 (October 1971), pp. 2647–62.
10. V. E. Beneš, "Models and Problems of Dynamic Memory Allocation," *Applied Probability—Computer Science: The Interface,* Vol. I, ed. R. L. Disney and T. J. Ott, Boston: Birkhauser, 1982, pp. 89–135.
11. E. G. Coffman, Jr., T. T. Kadota, and L. A. Shepp, "A Stochastic Model of Fragmentation in Dynamic Storage Allocation," SIAM J. Comput., *14,* No. 2 (May 1985), pp. 416–25.
12. G. F. Newell, "The M/M/∞ Service System With Ranked Servers in Heavy Traffic," Lecture Notes in Economics and Math. Systems, *231,* New York: Springer-Verlag, 1984.
13. L. A. Gimpelson, "Analysis of Mixtures of Wide and Narrow Band Traffic," IEEE Trans. Commun. Technol., *13* (1965), pp. 258–66.
14. E. Wolman, "The Camp-On Problem for Multiple-Address Traffic," B.S.T.J., *51,* No. 6 (July–August 1972), pp. 1363–422.
15. K. J. Omahen, "Capacity Bounds for Multiresource Queues," J. Assoc. Comput. Mach., *24,* No. 4 (October 1977), pp. 646–63.
16. E. Arthurs and J. S. Kaufman, "Sizing a Message Store Subject to Blocking Criteria," *Performance of Computer Systems,* ed. M. Arato, A. Butrimenko, and E. Gelenbe, Amsterdam: North-Holland, 1979, pp. 547–64.
17. L. Green, "A Queueing System in Which Customers Require a Random Number of Servers," Oper. Res., *28,* No. 6 (November–December 1980), pp. 1335–46.
18. P. H. Brill and L. Green, "Queues in Which Customers Receive Simultaneous Service From a Random Number of Servers: A System Point Approach," Manage. Sci., *30,* No. 1 (January 1984), pp. 51–68.
19. L. Green, "A Multiple Dispatch Queueing Model of Police Patrol Operations," Manage. Sci., *30,* No. 6 (June 1984), pp. 653–64.
20. A. Federgruen and L. Green, "An M/G/c Queue in Which the Number of Servers Required is Random," J. Appl. Probab., *21,* No. 3 (September 1984), pp. 583–601.
21. D. Mitra and P. J. Weinberger, "Probabilistic Models of Database Locking: Solutions, Computational Algorithms, and Asymptotics," J. Assoc. Comput. Mach., *31,* No. 4 (October 1984), pp. 855–78.
22. D. Mitra, unpublished work.
23. D. P. Heyman, "Asymptotic Marginal Independence in Large Networks of Loss Systems," Bell Communications Research, Holmdel, 1985. Presented at the ORSA/TIMS Applied Probability Conf., Williamsburg, Va., January 1985.
24. W. J. Hery, private communication.
25. J. T. Wittbold, AT&T Communications, private communication.
26. J. M. Akinpelu, "The Overload Performance of Engineered Networks With Nonhierarchial and Hierarachial Routing," AT&T Bell Lab. Tech. J., *63,* No. 7 (September 1984), pp. 1261–82.
27. D. L. Jagerman, "Some Properties of the Erlang Loss Functions," B.S.T.J., *53,* No. 3 (March 1974), pp. 525–51.
28. D. L. Jagerman, "Methods in Traffic Calculations," AT&T Bell Lab. Tech. J., *63,* No. 7 (September 1984), pp. 1283–310.
29. W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol. I, Third Edition, New York: Wiley, 1968.
30. D. R. Smith and W. Whitt, "Resource Sharing for Efficiency in Traffic Systems," B.S.T.J., *60,* No. 1 (January 1981), pp. 39–55.
31. W. Whitt, unpublished work.
32. D. J. Daley, "Stochastically Monotone Markov Chains," Zeitschrift Wahrscheinlichkeitstheorie Verw. Geb., *10* (1968), pp. 305–17.
33. J. Keilson and A. Kester, "Monotone Matrices and Monotone Markov Processes," Stoch. Proc. Appl., *5,* No. 3 (July 1977), pp. 231–41.
34. A. Kester, *Preservation of Cone Characterizing Properties of Markov Chains,* Ph.D. Thesis, University of Rochester, 1977.
35. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models,* ed. D. J. Daley, New York: Wiley, 1983.
36. W. A. Massey, "An Operator Analytic Approach to the Jackson Network," J. Appl. Probab., *2* (June 1984), pp. 379–93.
37. W. A. Massey, "Open Networks of Queues: Their Algebraic Structure and Estimating Their Transient Behavior," Adv. Appl. Probab., *16,* No. 1 (March 1984), pp. 176–201.
38. W. A. Massey, unpublished work.

39. N. Dunford and J. T. Schwartz, *Linear Operators, Part I: General Theory*, New York: Interscience, 1958.
40. W. Whitt, "Open and Closed Models for Networks of Queues," AT&T Bell Lab. Tech. J., *63*, No. 9 (November 1984), pp. 1911–79.
41. E. Fuchs and P. E. Jackson, "Estimates of Distributions of Random Variables for Certain Communications Models," Commun. ACM, *13*, No. 12 (December 1970), pp. 752–7.
42. P. F. Pawlita, "Traffic Measurements in Data Networks, Recent Measurement Results, and Some Implications," IEEE Trans. Commun., *COM-29*, No. 4 (April 1981), pp. 525–35.
43. W. T. Marshall and S. P. Morgan, unpublished work.
44. R. Schassberger, "Insensitivity of Steady-State Distributions of Generalized Semi-Markov Processes With Speeds," Adv. Appl. Probab., *10*, No. 4 (December 1978), pp. 836–51.
45. D. Y. Burman, "Insensitivity in Queueing Systems," Adv. Appl. Probab., *13*, No. 4 (December 1981), pp. 846–59.
46. W. Whitt, "Continuity of Generalized Semi-Markov Processes," Math. Oper. Res., *5*, No. 4 (November 1980), pp. 494–501.
47. E. Brockmeyer, H. L. Halstrom, and A. Jensen (eds.), *The Life and Works of A. K. Erlang*, Copenhagen: Danish Academy of Sciences, 1948.
48. F. Baskett et al., "Open, Closed and Mixed Networks of Queues With Different Classes of Customers," J. Assoc. Comput. Mach., *22*, No. 2 (April 1975), pp. 248–60.
49. F. P. Kelly, "Networks of Queues," Adv. Appl. Prob., *8*, No. 2 (June 1976), pp. 416–23.
50. R. Schassberger, "The Insensitivity of Stationary Probabilities in Networks of Queues," Adv. Appl. Probab., *10*, No. 4 (December 1978), pp. 906–12.
51. K. Matthes, "Zur Theorie der Bedienungsprozesse," Trans. Third Prague Conf. Inf. Theory, Prague, 1962.
52. A. D. Barbour, "Networks of Queues and the Method of Stages," Adv. Appl. Probab., *8*, No. 3 (September 1976), pp. 584–91.
53. S. S. Lam, "Queueing Networks With Population Size Constraints," IBM J. Res. Develop., *21*, No. 4 (July 1977), pp. 370–8.
54. P. Franken et al., *Queues and Point Processes*, Berlin: Akademie-Verlag, 1981.
55. A. E. Eckberg, "Generalized Peakedness of Teletraffic Processes," Proc. Tenth Int. Teletraffic Congress, Montreal, June 1983, p. 4.4 b.3.
56. A. A. Fredericks, "Approximating Parcel Blocking via State Dependent Birth Rates," Proc. Tenth Int. Teletraffic Congress, Montreal, June 1983, p. 5.3.2.
57. W. Whitt, "Heavy-Traffic Approximations for Service Systems With Blocking," AT&T Bell Lab. Tech. J., *63*, No. 5 (May–June 1984), pp. 689–708.
58. R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing*, New York: Holt, Rinehart and Winston, 1975.
59. S. Karlin and Y. Rinott, "Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Totally Positive Distributions," J. Multivar. Anal., *10*, No. 4 (December 1980), pp. 467–98.
60. T. Kamae, U. Krengel, and G. L. O'Brien, "Stochastic Inequalities on Partially Ordered Space," Ann. Probab., *5*, No. 6 (December 1977), pp. 899–912.
61. P. Billingsley, *Convergence of Probability Measures*, New York: Wiley, 1968.
62. W. Whitt, "On the Heavy-Traffic Limit Theorem for GI/G/∞ Queues," Adv. Appl. Probab., *14*, No. 1 (March 1982), pp. 171–90.
63. D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, New York: Springer-Verlag, 1979.
64. T. Lindvall, "Weak Convergence of Probability Measures and Random Functions in the Function Space $D[0, \infty)$," J. Appl. Probab., *1* (March 1973), pp. 109–21.
65. W. Whitt, "Some Useful Functions for Functional Limit Theorems," Math. Oper. Res., *5*, No. 1 (February 1980), pp. 67–85.
66. L. Arnold, *Stochastic Differential Equations: Theory and Applications*, New York: Wiley, 1974.
67. W. Whitt, "Weak Convergence of Probability Measures on the Function Space $C[0, \infty)$," Ann. Math. Statist., *41*, No. 3 (June 1970), pp. 939–44.
68. K. R. Parthasarathy, *Probability Measures on Metric Spaces*, New York: Academic Press, 1967.
69. W. Whitt, "Comparing Counting Processes and Queues," Adv. Appl. Probab., *13*, No. 1 (March 1981), pp. 207–20.
70. E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, New York: Wiley, 1966.

71. F. P. Kelly, "Blocking Probabilities in Large Circuit-Switched Networks," Statistical Laboratory, University of Cambridge, England, 1985.
72. I. B. Ziedens and F. P. Kelly, "Loss Probabilities in Circuit-Switched Star Networks," Statistical Laboratory, University of Cambridge, England, 1985.
73. D. Mitra, unpublished work.
74. P. M. Lin et al., "Analysis of Circuit-Switched Networks Employing Originating Office Control With Spill Forward," IEEE Trans. Commun., *COM-26*, No. 6 (June 1978), pp. 754–65.
75. A. Girard and Y. Ouimet, "End-to-End Blocking for Circuit-Switched Networks: Polynomial Algorithms for Some Special Cases," IEEE Trans. Commun., *COM-31*, No. 12 (December 1983), pp. 1269–73.
76. G. Iazolla, P. J. Courtois, and A. Hordijk, *Mathematical Computer Performance and Reliability*, Amsterdam: North-Holland, 1984.

AUTHOR

Ward Whitt, A.B. (Mathematics), 1964, Dartmouth College; Ph.D. (Operations Research), 1968, Cornell University; Stanford University, 1968–1969; Yale University, 1969–1977; AT&T Bell Laboratories, 1977—. At Yale University, from 1973–1977, Mr. Whitt was Associate Professor in the departments of Administrative Sciences and Statistics. At AT&T Bell Laboratories he is in the Operations Research Department of the Systems Analysis Center, where the primary mission is to investigate and improve the product realization process.