AT&T Bell Laboratories Technical Journal
Vol. 63, No. 9, November 1984
Printed in U.S.A.

Open and Closed Models for Networks of Queues

By W. WHITT*

(Manuscript received April 9, 1984)

This paper investigates the relationship between open and closed models for networks of queues. In open models, jobs enter the network from outside, receive service at one or more service centers, and then depart. In closed models, jobs neither enter nor leave the network; instead, a fixed number of jobs circulate within the network. Open models are analytically more tractable, but closed models often seem more realistic. Hence, this paper investigates ways to use open models to approximate closed models. One approach is to use open models with specified expected equilibrium populations. This fixed-population-mean method is especially effective for approximately solving large closed models, where "large" may mean many nodes or many jobs. The success of these approximations is partly explained by limit theorems: Under appropriate conditions, the fixed-population-mean method is asymptotically correct. In some cases the open-model methods also yield bounds for the performance measures in the closed models.

I. INTRODUCTION AND SUMMARY

Queueing network models are now widely used to analyze communication, computing, and production systems. A relatively well-developed theory exists for the Markov Jackson network models and various extensions that have a product-form equilibrium (steady-state) distribution.[1-6] In this paper we consider both the product-form models for which exact solutions are possible and more complicated nonproduct-

* AT&T Bell Laboratories.

Copyright © 1984 AT&T. Photo reproduction for noncommercial use is permitted without payment of royalty provided that each reproduction is done without alteration and that the Journal reference and copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free by computer-based and other information-service systems without further permission. Permission to reproduce or republish any other portion of this paper must be obtained from the Editor.

form models for which approximations are needed. We are primarily motivated by the desire to develop new approximations for non-product-form closed models. We discuss methods for modifying the Queueing Network Analyzer (QNA) software package[7] so that it can be used to calculate approximate congestion measures for closed models as well as open models. Our general approach to the closed models is to apply previous techniques for open models. Hence, we investigate the relationship between open and closed models.

For simplicity, we first consider a Markov Jackson network model with First-Come First-Served (FCFS) multiserver nodes (service centers) and one job class. Flow within the network is determined by stochastic routing probabilities: Each job completing service at node i goes immediately to node j with probability q_{ij}, independent of the history of the system. The individual service rates and external arrival rates, if any, are independent of the state. The service-time distributions are exponential and the external arrival processes, if any, are Poisson. It will be clear that the ideas generalize.

1.1 Open and closed models

These models can be classified as open or closed. In an open model, jobs enter the network at random from outside at a fixed rate, receive service at one or more nodes, and eventually leave the network. Thus, with an open model the total external arrival rate or throughput is an independent variable (specified as part of the model data), and the number of jobs in the system is a dependent variable (whose equilibrium distribution is described in the model solution). On the other hand, in a closed model there is a fixed population of jobs in the network. Hence, with a closed model the number of jobs in the system is an independent variable (specified as part of the model data), and the throughput (which may be defined, for example, as the departure rate from some designated node) is a dependent variable (to be calculated and described in the model solution). Since the individual service rate is part of the model data, knowing the throughput is equivalent to knowing the utilization, which is the expected proportion of the servers at the designated node that are busy in equilibrium.

Of course, there also are more complicated models, in which the simple dichotomy above is not valid. For example, Jackson introduced models in which the external arrival rate can depend on the total number of jobs in the network.[1] Then neither the external arrival rate nor the network population is fixed. There are also mixed models, which have some classes with fixed populations and other classes with fixed external arrival rates.[3,6] We will not consider these more general models, but we note that open models can be used to approximate mixed models in the same way that they can be used to approximate closed models.

It might seem that open models would be more appropriate for most applications because jobs do usually come from outside, flow through the system, and eventually depart. However, closed models are often used instead. The representation of flow through the system, i.e., the throughput, is easily handled in a closed model by assuming that a new job enters the system to replace an old one whenever the old one has received all of its required service. This can be represented in the closed model by a transition to a designated exit-entry node. At this node, arriving jobs complete all of their required service, and departures are new jobs. The rate of transitions through this node (which is both the arrival rate and the departure rate) can be regarded as the throughput. If no such exit-entry node exists originally, it is easy to add such a node. The modified network with the additional node is equivalent to the original network if all jobs at this new exit-entry node have zero service time.

Evidently, closed models are often applied because it seems natural to regard the number of jobs in the system as the independent variable and the throughput as the dependent variable. The number of jobs in the system is often subject to control; the queueing analysis is desired to determine the associated throughputs and response times. For example, in production systems, new jobs usually do not arrive at random; they are scheduled. In fact, this view was the main reason that Jackson extended queueing network theory to cover closed models:[1]

> This extension of the author's earlier work is motivated by the observation that real production systems are usually subject to influences which make for increased stability by tending, as the amount of work-in-process grows, to reduce the rate at which new work is injected or to increase the rate at which processing takes place.

Similarly, in computing systems the total number of jobs in device queues tends to be limited by resource constraints, so that it is natural to specify the number of jobs (the multiprogramming level) as a decision variable and then calculate the associated throughput (see p. 116 of Ref. 6). Also, in time-sharing systems the number of jobs is limited by the number of sources (terminals), so that the total number of jobs is not unbounded (see p. 60 of Ref. 6). Hence, even though closed models are significantly more difficult to analyze because of the normalization constant or partition function, there are good reasons for applying them.

1.2 The fixed-population-mean method

In this paper we propose and investigate a different approach that may sometimes be an attractive alternative. We propose using the open model with specified expected equilibrium population, which we

refer to as the Fixed-Population-Mean (FPM) method. With the FPM method, we have the analytically more elementary open model, but the number of jobs (or, more precisely, its mean) becomes an independent variable and the throughput or total arrival rate becomes a dependent variable. Even though some of the initial modeling assumptions may not seem appropriate (e.g., unlimited population and Poisson arrivals), we believe that the approach has potential. With regard to the modeling assumptions, it is important to remember that the model solutions are describing only the equilibrium or, equivalently, long-run averages. Moreover, the closed model assumptions are often not entirely appropriate either. In many situations where closed models are applied, the total population is not nearly fixed. The FPM equilibrium solution may better describe these systems. Moreover, we can modify the FPM method in various ways to obtain a better description.

Even when a closed model is deemed appropriate, the open model with the FPM method can be useful because if often provides a convenient approximation for the more difficult closed model. Certainly the required computation is significantly reduced. In some cases, throughputs can be calculated by hand by the FPM method when some computer codes for closed models are unable to obtain any solution. Moreover, in many cases the results are very close.

The FPM method also forms the basis for one procedure to calculate approximate congestion measures for closed non-Markov networks containing multiserver FCFS nodes with nonexponential service-time distributions, using previously developed approximation procedures for open non-Markov networks such as the Queueing Network Analyzer (QNA).[7] In fact, the primary motivation for this work was the desire to modify QNA so that it can analyze closed models as well as open models. The FPM method is one way to do this. Several possible approaches for calculating approximate congestion measures for non-Markov closed networks are described in Section X.

For the basic Jackson network we are now considering, we implement the FPM method by identifying the external arrival rate that yields the specified expected equilibrium number of jobs in the system. The standard application would be to a system that was previously analyzed by a closed model. Consider such a closed model with a designated exit-entry node. In the closed model, an arrival to this node from elsewhere in the network completes its required service, and a departure represents a new job entering the system. To obtain the associated open model, cut the flow into this exit-entry node, let all internal arrivals into this node leave the system, and insert an external Poisson arrival process. The FPM throughput is the rate or intensity of the external Poisson arrival process for which the expected equilib-

rium population has the specified value. The approximate mean number of jobs at each node is the mean number in the open model with the FPM arrival rate.

In fact, in the FPM procedure described above, it is not necessary to have a special exit-entry node. Any node can serve as the exit-entry node. For the Jackson product-form network we are considering, the equilibrium distribution of the resulting open model is independent of the node chosen. Choosing the exit-entry node can be important, however, if we do not have a product-form network. Then it may also be appropriate to let the new external arrival process be something other than a Poisson process.

In this introductory section we give a few elementary examples to illustrate the FPM method. However, the primary motivation is the need for approximate methods to analyze more complicated models, e.g., with multiple job classes, nonexponential FCFS servers, priorities, etc. It should be clear that the FPM method is a general approach that applies to these more complicated models. We believe that the performance of the FPM method for Jackson networks indicates the performance that can be expected for more complicated models.

Example 1. Consider a closed Markov cyclic network of single-server FCFS queues with K jobs of a single class. Let there be n_1 nodes having mean service time 1 and n_2 nodes having mean service time τ, arranged in any order. As usual, cyclic means that all departures from node j go next to node $j+1$ for $1 \leq j \leq n_1 + n_2 - 1$ and all departures from node $n_1 + n_2$ go next to node 1. To apply the FPM method, cut the flow into one node, let all original arrivals on that arc leave the system, and insert an external Poisson arrival process. We identify the external arrival rate in the associated open model, say λ, such that the expected equilibrium total population is K. Since we have a cyclic network, the arrival rate at each node is the external arrival rate. (Otherwise, we would have to solve the traffic rate equations.) Recalling that the equilibrium distribution in the open model is equivalent to independent M/M/1 queues, we solve

$$\frac{n_1\lambda}{1-\lambda} + \frac{n_2\lambda\tau}{1-\lambda\tau} = K$$

for λ, which is a quadratic equation. (If there were m different service rates, then we would have a polynomial of degree m.)

To illustrate, if $K = 20$, $n_1 = n_2 = 10$ and $\tau = 1.2$, then the approximate throughput is = 0.45. In Section IV we prove that this is a lower bound for the throughput in the original closed model. In (15) we suggest as a possible improvement $\lambda(n_1 + n_2 + K)/(n_1 + n_2 + K - 1)$, which in this case is 0.46. The actual throughput in the original closed network also turns out to be 0.46. This is easily determined using any

software package for closed Markovian networks of queues; we used PANACEA.[8]

Since the mean service times at different nodes do not differ much, we could also use a quicker approximation, based on a linear equation instead of the quadratic equation, obtained by assuming that all $n_1 + n_2$ nodes have mean service time $\bar{\tau} = (n_1 + n_2\tau)/(n_1 + n_2)$, and then applying the FPM method, which yields the almost instantaneous approximation $K\bar{\tau}/(K + n_1 + n_2) = 20/44 = 0.45$ for the throughput.

This last balanced network approximation can also be used directly in the closed network, which corresponds to combining the last two suggested improvements. (See Section III.) The resulting approximate throughput by this method is $K\bar{\tau}/(K + n_1 + n_2 - 1) = 0.47$. This direct balanced network approximation for closed networks is in fact an upper bound on the throughput in the closed model, as was first shown by Zahorjan et al.[9] (also see Refs. 10 and 11).

If we use $\lambda = 0.45$ as the approximate throughput by the FPM method, then with the M/M/1 formula the mean number of jobs at each node with mean service time 1 (1.2) is 0.82 (1.18); for the closed model, it is 0.83 (1.17). The expected sojourn time at each of these nodes by the FPM method is 1.82 (2.62); for the closed model, it is 1.79 (2.53). For practical purposes, the standard congestion measures calculated by the FPM method agree with those for the closed model in this example.

Note that the FPM solution does not change if we multiply the population and the number of nodes of each type by a common constant. It turns out that the quality of the approximation improves as the network grows in this way. On the one hand, this means that the FPM method does not perform well when there are few nodes, e.g., when $n_1 = n_2 = 1$ here. On the other hand, the FPM method tends to perform well for the large models that are more difficult for closed network algorithms. In fact, in Section V we prove that the FPM method is asymptotically correct for such growing closed networks. This asymptotic property of large closed networks was apparently first observed by Gordon and Newell.[2] We contribute by providing a rigorous proof based on the local central limit theorem for sums of independent and identically distributed random vectors.[12] Also, we stress the significance of the FPM method in this asymptotic analysis. Algorithms for closed models have difficulty as the number of nodes increases. Evidently, no existing closed-network algorithm is able to handle the case of 200 nodes and 200 customers for this numerical example. With the aid of new asymptotic theory,[13,14] PANACEA[8] is able to solve much larger networks, but the asymptotic theory does not apply to this example because it requires a decoupling infinite-server node (see Section 1.4).

In general, when we apply the FPM method, we do not get a quadratic equation. However, the expected equilibrium population in the newly created open network is an increasing function of the external arrival rate, so that it is not difficult to identify the external arrival rate which yields the desired fixed population mean by a search procedure. In fact, it is usually possible to quickly obtain the desired throughput with a programmable hand calculator that can find the roots of an equation. (At the expense of some added complexity, this same general approach can be used for multiple job classes. We give an efficient iterative algorithm for special cases involving infinite-server nodes in Section VI.)

However, it is usually not necessary to carry out such a special inversion procedure. As is standard for closed models, we usually want to determine the throughput as a function of the (expected) network population. Hence, we simply solve the open model for a range of possible external arrival rates and express the expected equilibrium population as a function of the external arrival rate. It is then easy to invert the function if desired. Moreover, we can also describe the resulting population variability in the open model as a function of the external arrival rate. Thus, the FPM method consists of little more than using open models in situations where closed models were used before. Our object, then, is to better understand the relationships between these two kinds of models. We propose some algorithms and obtain some insight about when they will work well and when they will not.

When both closed and open models are available, the appropriate model might be chosen according to which better describes the population variability. We suggest using estimates of the population variance to help identify an appropriate model. It turns out that the population variability in the open model is often less than might be expected (Section II), so that the two models are often remarkably similar. For larger networks (large population or many nodes), the differences are often small relative to the quality of data typically available for modeling fitting. When this is the case, the open model is usually preferred because it is much easier to analyze.

The possible advantage of closed models over open models is also reduced if we do not restrict attention to Markov open models. For example, if we use QNA to approximately analyze a non-Markov open model, then we have an additional degree of freedom in modeling the variability because we can select variability parameters for each service-time distribution and each arrival process. If the actual arrivals are scheduled, as in many production systems, then it is natural to use clocked arrivals in QNA, i.e., deterministic interarrival times, which is achieved by setting the variability parameter for the external arrival

process equal to zero. With an open model, we are not forced to have a Poisson external arrival process. From direct modeling considerations, the open model with clocked external arrival processes is often more realistic than the closed model.

Furthermore, given that we are actually interested in a closed model, the variability parameters offer the possibility of improved approximations by open models. Since the population constraint in the closed model tends to reduce the variability (see Section IV), a promising heuristic approximation procedure with the FPM method (suggested by H. Heffes) is to reduce the variability parameters in the approximating open model. For example, with a Jackson network of single-server queues, we might treat each node as a D/M/1 queue instead of an M/M/1 queue, but the actual procedure would have to be more sophisticated. The general approach using QNA for non-Markov closed models is to cut the flow into one node and replace it with an external arrival process. First, as described in Section X, we can let the variability parameter of the external arrival process be such that it agrees with the variability parameter of the departure process from the network. We use QNA to calculate approximate variability parameters for the arrival process to each node. Afterwards, to improve the approximation of the closed model, we can systematically reduce all these variability parameters. The reduction should depend on the network parameters, with the variability parameters evidently being reduced less as the number of nodes or the number of jobs increases. We briefly investigate this possibility, but we have just begun studying refined approximation procedures of this kind.

1.3 The finite-waiting-room refinement

We also propose a refinement of the FPM method for approximating closed models, which is especially useful for small models. We apply the network population constraint given for the closed model to each node separately in the open model. When there are K jobs in the closed model, we allow at most K jobs at each node in the open model. However, we implement this Finite-Waiting-Room (FWR) approximation within the product-form equilibrium distribution of the open model. We act as if there is capacity K at each node in the open model, but we do not analyze the modified open model exactly. Instead, we keep the product-form equilibrium distribution in the open model, and modify the distribution of the number of jobs at each node.

If N_i^o is the equilibrium number of jobs at node i in the open model without the refinement, then we use the conditional distribution of N_i^o given that $N_i^o \leq K$. Since N_i^o has the distribution of a birth-and-death process in a Jackson network, this conditional distribution obtained simply by truncating the original distribution at K and

renormalizing is tantamount to imposing a finite waiting room at that node in isolation. [This conditioning can also be used as an approximation for more general models (see Refs. 15 and 16).]

Let $\bar{N}_i^o$ be the equilibrium number of jobs at node i by the FWR method; then

$$P(\bar{N}_i^o = k) = P(N_i^o = k \mid N_i^o \leq K) = P(N_i^o = k)/P(N_i^o \leq K). \quad (1)$$

For an M/M/1 queue, the mean is $EN_i^o = \rho_i/(1 - \rho_i)$ and the utilization is $u_i^o = P(N_i^o > 0) = \rho_i$, where $\rho_i = \lambda_i/\mu_i$ is the traffic intensity at node i, based on the net arrival rate λ_i and service rate μ_i. The corresponding quantities $E\bar{N}_i^o$ and $\bar{u}_i^o$ with the FWR method are

$$\begin{aligned} E\bar{N}_i^o &= w(\rho_i, K)EN_i^o \\ w(\rho_i, K) &= (1 - (K+1)\rho_i^K + K\rho_i^{K+1})/(1 - \rho_i^{K+1}) \\ \bar{u}_i^o &= P(\bar{N}_i^o > 0) = (\rho_i - \rho_i^{K+1})/(1 - \rho_i^{K+1}), \end{aligned} \quad (2)$$

provided that $\rho_i \neq 1$ (see Section 2.5 of Ref. 17). Obviously, $E\bar{N}_i^o < EN_i^o$ and $\bar{u}_i^o < u_i^o$. If we let $\bar{u}_i^o$ be free, fix u_i^o, and let $E\bar{N}_i^o = EN_i^o$, then $\bar{u}_i^o > u_i^o$ (see Section IV), which is a refinement in the right direction. Moreover, if u_i^c is the utilization of node i and N_i^c is the equilibrium number of jobs at node i in the closed model, then $\bar{u}_i^o \leq u_i^c$ when $E\bar{N}_i^o \leq EN_i^c$. Since the ratio of the utilizations at any two nodes in the closed model is the same as in the open model,[5,6] we thus obtain a valid lower bound on u_i^c by this procedure, namely,

$$u_i^c \geq \bar{u}_i^* \equiv \min_j \{(u_i^o/u_j^o)\bar{u}_j^o\}. \quad (3)$$

We need (3) to obtain the valid lower bound because the property $u_i^o/u_j^o = u_i^c/u_j^c$ for all i and j does not hold for $\bar{u}_i^o/\bar{u}_j^o$. [See (15) and Section IV.]

Example 2. Consider a closed Markov network with n single-server nodes and K jobs. Let the service rates and net arrival rates be identical, so that the equilibrium distribution is symmetric. When $n = 4$ and $K = 2$, the server utilizations by direct analysis of the closed model, the FPM method, and the FPM/FWR method are, respectively, 0.400, 0.333, and 0.384. By using the FWR refinement, the error is reduced from 42 percent to 4 percent.

Using the FWR refinement to the FPM method can yield external arrival rates for which $\rho_i \geq 1$ at some nodes. Limited numerical experience indicates that the quality of the approximation often deteriorates in this case.

1.4 Decoupling infinite-server nodes

There need not be many nodes for the FPM method to be effective. The FPM method is particularly appealing to approximately solve

closed Markovian networks with few nodes but a large population and an Infinite-Server (IS) node with relatively low service rate: For these models, the FPM method extends easily to multiple job classes. However, the need for help with these difficult models is much less now because an efficient algorithm for them has recently been developed by McKenna, Mitra, and Ramakrishnan,[13,14] which is implemented in their PANACEA software package.[8]

The PANACEA algorithm exploits integral representations and asymptotic expansions to reduce the original large closed network to many much smaller closed networks. Under appropriate conditions, the difficult partition function of the original closed model and related quantities such as the utilization of a particular class at a particular node can be represented by asymptotic expansions in which the coefficients are constructed from the partition functions of the smaller closed networks (the pseudonetworks in Ref. 14, which typically involve at most three classes and a total of seven customers). Moreover, the asymptotic expansions permit a thorough analysis of the truncation errors: The truncation error is less in absolute value than the first neglected term and has the same sign.

The asymptotic expansions underlying the new capabilities in PANACEA are based on several assumptions. First, it is assumed that each class visits an IS node [see (27) of Ref. 14]. Second, it is assumed that the population of each class is large [see (17) of Ref. 14]. Third, it is assumed that the individual service rates at the IS node are significantly lower than the service rates in the rest of the network [see (18) of Ref. 14]. Finally, it is assumed that utilizations of the non-IS nodes are not close to their critical values, i.e., they are not in heavy traffic [see (29) through (31) of Ref. 14]. It is worth noting that these assumptions are often realistic—e.g., in computing systems where the IS nodes correspond to "think times" at terminals.

It turns out that the FPM method tends to work well under these same conditions. Unlike PANACEA, however, the FPM method is an approximation. (The asymptotic expansions in PANACEA also can be regarded as approximations, but of a different kind; they are a numerical method that can achieve any degree of accuracy given enough computation. On the other hand, the FPM method changes the model, so that the answers are good only if the two model solutions are close.) The FPM method in this situation can be derived by a procedure that at first seems to be different from the FPM method. This alternate procedure is motivated by the observation that under the stated conditions the departure processes from the special IS nodes tend to behave much like Poisson processes. Moreover, the subnetwork without the IS nodes tends to behave much like an open network with an external Poisson arrival process. This is partly substantiated by

previous work[18] in which we showed that, under appropriate conditions, the departure process from an IS node with a fixed general stationary arrival process and a general service-time distribution approaches a Poisson process that is independent of the arrival process as the individual service rate at the IS node decreases. Reference 18 does not directly apply here because the arrival processes at the IS node are changing too, but Ref. 18 suggests that the departure processes from the IS nodes are approximately Poisson processes that are independent of the rest of the network under the stated assumptions. Corresponding limit theorems for the situation here are contained in Section VIII.

The key to our procedure for these models, as with the asymptotic expansions underlying PANACEA, is a large population and the presence of IS nodes with relatively low service rates. The FPM method can be used more generally, but there is stronger supporting logic with the IS nodes. The FPM method works well if there are several IS nodes, as long as each class visits one of them, but for simplicity we assume that there is a single IS node visited by all classes. Also, each class can visit more than one IS node, but we assume that only one IS node has relatively low service rate, so that jobs tend to accumulate there. We use this IS node to decouple the network. We let its departure process leave the system and replace it by an external arrival process. The external arrival process is a set of independent Poisson processes, with one Poisson process for each class. Equivalently, there is a simple Poisson external arrival process and fixed probabilities that each arrival belongs to one of the classes. We approximately solve the original closed network by identifying the appropriate external arrival rates for the associated open model. We use the special IS node to determine what rates are appropriate. We do this by simply equating the arrival and the departure rates for each class at the IS node. It turns out that this procedure is equivalent to the FPM method discussed above (see Section VI).

Example 3. To illustrate the FPM method with an IS node having relatively low service rate, we consider a central processor model treated by McKenna, Mitra, and Ramakrishnan.[13] This is a closed cyclic network with only two nodes. The first node is the CPU, which has a single server, where service is provided according to the processor-sharing discipline. The second node is a "think" node, which is an IS node, representing independent delays at terminals before a job is next sent to the CPU. (Because of insensitivity properties,[5,6,19,20] only the means of the service-time distributions matter for the equilibrium distribution. We can also equivalently regard the CPU as an FCFS node with an exponential service-time distribution.) We shall consider the case of one job class, which is test problem I described in Table I

Table I—A comparison of throughputs using closed and open models for the two-node single-class network model of a central processor in Example 3 of Section 1.4

	Throughput or Utilization of CPU							
	For the Closed Model			For the Open Model Using FPM Method				
	From Ref. 13						Taylor's Series (39)	
Number of Jobs	By CADS	By PANACEA	Version 2.1 of PANACEA	First Upper Bound (30)	First Lower Bound (31)	FPM Solution (28) and (38)	Two Terms	Three Terms
10	0.0417	0.0414	0.0415	0.0417	0.0415	0.0415	0.0415	0.0415
50	0.207	0.207	0.2073	0.2083	0.2072	0.2072	0.2074	0.2072
100	Breakdown	0.413	0.4138	0.4167	0.4137	0.4137	0.4150	0.4143
200	Breakdown	0.819	0.8204	0.8333	0.8123	0.8150	0.8298	0.8269

of Ref. 13. The mean service times at the two nodes are 1 and 240, respectively, so indeed the IS node has relatively low individual service rate.

This example was significant in Ref. 13 because it demonstrated the advantage of PANACEA over previous closed network algorithms, in particular CADS.[21] In several cases CADS was unable to obtain a solution. This example is also significant here because for it the FPM method is both easy and accurate. The FPM method only requires the solution of a quadratic equation [see (38) in Section VII]. A quick upper bound on the CPU utilization can be found by simply multiplying the population times the IS individual service rate [see (30) in Section VI]. Both of these methods perform remarkably well. The throughput results for several population sizes are described in Table I. The CPU node has a processor-sharing service discipline with mean processing time of $\mu_1^{-1} = 1$. Hence, the throughput of jobs at the CPU is equal to the utilization of the CPU. Node 2 is an infinite-server delay node representing the think time of users at terminals. The mean think time is $\mu_2^{-1} = 240$.

For the cases in Table I it is apparent that even the trivial upper bound is adequate for practical purposes. Moreover, as we have remarked, the FPM throughput itself is a lower bound (see Section IV), so that from the FPM method alone we can determine that the quality of the approximation is satisfactory. Since the FPM throughput is a lower bound, from Table I we see that in some cases the FPM throughput is actually slightly more accurate than the published values in Ref. 13, but of course the differences are not significant for practical purposes. In the difficult case of 200 jobs, Version 2.1 of PANACEA terminated with lower and upper bounds of 0.8129 and 0.8204, based on four terms of the asymptotic expansion. In the other cases the two bounds coincide for the specified accuracy. The main point is that essentially the same answers can be obtained quickly by hand. (See Sections VI and VII for additional discussion.)

With the FPM method we avoid closed networks and the associated partition functions entirely. Instead, we approximately solve the original closed network by solving a related open model. In the case of multiple job classes, we iteratively solve a sequence of associated open networks. By working with open networks, we never calculate the complete distribution of the number of jobs of each class at each node. With open networks it suffices to work with expected values. By exploiting simple monotonicity properties, we are also able to give upper and lower bounds on the desired approximate solution at each iteration. Finally, we are able to treat very general networks; e.g., the subnetwork can have multiserver nodes. In fact, the approximation procedure is ideal for the closed-model analogs of the non-Markov

networks analyzed by QNA,[7] in which there are FCFS nodes with nonexponential service-time distributions. For the first step of the analysis, in which we replace the departure process from the IS node by an external Poisson arrival process, the FPM method is still asymptotically correct. Hence, for these closed non-Markov models with decoupling IS nodes, it appears that the FPM method with QNA should perform about as well as the original QNA approximation for the corresponding open non-Markov model.

1.5 The rest of this paper

The rest of this paper is organized as follows. In Section II we discuss population variability in open networks, and show that it tends to be relatively small in large networks, which supports using the FPM method to approximate large closed models. We also suggest measuring population variability to help decide which model to use. In Section III we compare the throughputs in closed models and open models (using the FPM method) for a special class of balanced networks, and we show that the differences are, for practical purposes, negligible when the network is large (but not when the network is small). This example provides a convenient, simple, quantitative characterization of the difference between the models, as far as throughputs are concerned. The explicit balanced network results also suggest possible refinements of the FPM method for unbalanced networks.

In Section IV we present theoretical results about the closed network and the FPM approximation. We prove that the utilizations and throughput calculated by the FPM method are always lower bounds for the corresponding quantities in the closed network (see Theorem 1). For the special case of single-server and infinite-server nodes, this result can also be deduced from Zahorjan.[22] We not only treat general nodes such as multiserver nodes, but we treat models with several job classes. To make our comparison and establish other properties of the closed model, we exploit the log-concavity[23] of the distribution of the number of jobs at each node in the associated open model. In various ways we show that the distribution of jobs is less variable in the closed network than in the associated open network. In particular, given ordered means at any node, we establish increasing concave stochastic order (see Theorem 2).[24] To do this, we introduce and apply the notion of one distribution being log-concave relative to another (see Definition 1).

It is intuitively clear that the population constraint should introduce negative dependence among the queue lengths at the different nodes. In Section IV we also show that recently developed concepts of multivariate negative dependence, such as reverse-rule distributions[25,26] and negative association,[27] are ideally suited to make this

idea precise. Indeed, the multivariate distribution of jobs at the different nodes has all of these properties. The closed Markovian network model can be regarded as a canonical example of negative dependence.

In Section V we present some additional theory to show that the FPM method tends to perform well for large networks. We prove that the FPM method is asymptotically correct as the number of nodes in the closed model increases with the number of customers per node held fixed (see Theorem 7). For this result, we can let the network grow in a cyclic manner so that the connectivity does not increase. To obtain a rigorous proof, we apply the local central limit theorem for partical sums of random vectors.[12] This result applies to Example 1 as a special case.

In Section VI we present the variant of FPM method to approximate closed networks with a large population and a decoupling IS node with relatively low service rate. We exploit monotonicity to obtain an efficient algorithm for multiple job classes. In Section VII we illustrate the FPM method in this context by considering the central processor model in Example 3. We consider cases involving two job classes as well as one. This example is taken from Ref. 13, so that we can conveniently compare the FPM method to numerical results for PANACEA and the CADS algorithm for closed models.[8,21]

In Section VIII we present theoretical results to support the FPM method in the context of Sections VI and VII. We show that the vector-valued queue-length process in the subnetwork of the closed model without the IS node converges in distribution to the corresponding stochastic process in the approximating open model with a Poisson external arrival process as the populations increase and the individual service rates at the IS node decrease appropriately. We establish convergence in distribution (weak convergence[28,29]) of both the stochastic processes (see Theorem 8) and the equilibrium distributions (see Theorem 9). Convergence of the departure process from the IS node to a Poisson process is established as in Refs. 18 and 30; convergence of the associated vector-valued queue-length process is established by model continuity.[31,32] As a consequence, the FPM method is asymptotically correct for the closed model under these conditions (see Theorem 12). In Section VIII we also make stochastic comparisons between the stochastic processes in the open and closed models, exploiting couplings or almost-surely ordered constructions as in Refs. 33 and 34. The stochastic comparisons are interesting in their own right, but they also play a role in establishing the convergence. We show that a first upper bound for the FPM method is also an upper bound for the closed model in terms of both transient and equilibrium throughputs and queue lengths (see Corollaries to Theorems 10 and 11).

In Section IX we discuss similar approximations for closed networks with a bottleneck node which is not an IS code. We propose a different approximation for closed networks with a bottleneck node. We delete the bottleneck node from the closed network, but we do not use the FPM method. Instead, the proposed approximation is to simply use the open model obtained by deleting the bottleneck node and replacing its departure process by an external arrival process generated by the service times at the bottleneck node. The difference between the total population and the expected population in the open subnetwork is the suggested approximation for the expected population at the bottleneck node. This approximation method is also asymptotically correct as the population grows. The vector-valued queue length process in the subnetwork of the closed network without the bottleneck node converges in distribution to the corresponding process in the open network as the population grows. This phenomenon is of course quite well known,[35,36] but some of the supporting theory here seems to be new.

In Section X we discuss methods for approximately solving non-Markov closed models. We indicate how existing procedures for non-Markov open networks such as the Queueing Network Analyzer (QNA) can be modified for this purpose. In particular, we describe in detail the changes in Ref. 7 to implement the FPM method.

In Section XI we provide some additional motivation for considering special algorithms to analyze non-Markov closed networks. It is sometimes claimed that Markov models with exponential service-time distributions adequately describe throughputs for single-server FCFS nodes with nonexponential service-time distributions with the same mean, but we show that this is not always the case. We use tight lower bounds on the throughput in closed models with FCFS single-server nodes and general service-time distributions identified by Arthurs and Stuck.[11] For highly variable distributions, the actual throughput can be much less than predicted by the Markov model. In fact, the Markov model can be arbitrarily bad. The true throughput can be arbitrarily close to zero, while the Markov model throughput is arbitrarily close to one.

In Section XII we make additional numerical comparisons that help put the different models and approximation procedures in perspective. We draw some conclusions in Section XIII.

This paper contains diverse material, ranging from heuristic algorithms and examples to theorems and proofs. These are intended to complement each other, but the primary algorithm sections (Sections VI and X) and mathematics sections (Sections IV, V, and VIII) can be read independently.

1.6 Other bounds and approximations

In this paper we introduce several approximation procedures and

establish several bounds for networks of queues. Of course, many other approximations and bounds have already been developed by others. In addition to the previously mentioned balanced network bounds and others in Refs. 9 through 11, there are useful bounds in Refs. 37 through 39. There is great potential for combining them in new ways. We focus on the basic Markov network models and natural non-Markov extensions obtained by allowing nonexponential service-time distributions and non-Poisson arrival processes. However, the results also have relevance for more complicated models and other approximation procedures.

For example, open-model representations such as the FPM method can be applied in conjunction with aggregation-decomposition approximation methods for closed networks, as suggested by Zahorjan.[22] The basic approach is to replace a subnetwork of a closed network by a single "composite" node with a state-dependent service rate (see pp. 165 through 172 of Ref. 6). For the product-form models, by Norton's theorem, the aggregation step is exact if when there are m_j jobs of class j in the subnetwork, the composite node service rate of class j is precisely the throughput rate for class j for the subnetwork in isolation (as a closed model with that population vector) (see pp. 100 and 106 of Ref. 6). The FPM method can be used as an approximation here to calculate approximate throughputs for the subnetworks.

The FPM method is a very natural idea, so no doubt it has been considered before. In fact, we have indicated that it appears in the asymptotic analysis in Ref. 2. The FPM method is also intimately related to another approximation procedure, which is called the Approximate Infinite Source (AIS) method and was proposed independently by Fredericks.[40] The idea in the AIS method is to replace a finite source by an infinite source, in particular a Poisson process, so that Little's formula[6,19] relating the throughput, expected population, and expected sojourn time remains valid. However, since the original population with a finite source is fixed, it is easy to see that this constraint is equivalent to making the expected equilibrium population in the open model coincide with the actual population in the closed model (with the finite source). Hence, aside from additional refinements, the AIS and FPM methods coincide. Fredericks illustrates the effectiveness of this approach with other examples, including a two-class priority service system with separate finite sources.

II. POPULATION VARIABILITY IN AN OPEN NETWORK

At first glance, it might seem that the open model with fixed mean number of jobs would always differ dramatically from the associated closed model with fixed actual number of jobs. It might seem that the population variability in the open network would necessarily be much

greater than in the closed network (for which there is no population variability at all). Indeed, for networks with few nodes there typically is a dramatic difference in the variability, but it turns out that the population variability of an open network tends to decrease as the number of nodes increases. This suggests that the FPM method should work well for large networks.

In one sense the variability in an open network increases as the number of nodes increases. Since the equilibrium numbers of jobs at the different nodes in an open model are independent, the variance of the equilibrium population is the sum of the variances at the nodes. So, roughly speaking, the population variance tends to grow as the network grows (assuming that the marginal distributions at the individual nodes do not change).

2.1 The population squared coefficient of variation

However, we believe that the squared coefficient of variation (the variance divided by the square of the mean) is usually a better measure of the relevant variability than the variance. It describes the variability relative to the mean. Suppose that in the open network under consideration there are n single-server nodes with the traffic intensity at node i being ρ_i. Since the equilibrium number, N_i^o, of jobs at node i has a geometric distribution, the mean, variance, and squared coefficient of variation of N_i^o are

$$E(N_i^o) = \rho_i/(1-\rho_i),$$
$$\mathrm{Var}(N_i^o) = \rho_i/(1-\rho_i)^2, \quad \text{and} \quad c^2(N_i^o) = 1/\rho_i.$$

Obviously, $c^2(N_i^o)$ can be arbitrarily large, so that we cannot expect the variability to be small with one single-server node.

The associated parameters of the equilibrium total number, N^o, of jobs in the entire network are

$$E(N^o) = E(N_1^o) + \cdots + E(N_n^o),$$
$$\mathrm{Var}(N^o) = \mathrm{Var}(N_1^o) + \cdots + \mathrm{Var}(N_n^o),$$
$$c^2(N^o) = \mathrm{Var}(N^o)/E(N^o)^2. \tag{4}$$

When the traffic intensities are all equal, i.e., $\rho_i = \rho$ for all i, $c^2(N^o) = 1/n\rho$, so that $c^2(N^o)$ tends to decrease rapidly as the number of nodes increases. There is a law of large numbers effect when there are many nodes.[41] This is also true as n increases when the traffic intensities are unequal, provided that $E(N_i^o)$ is asymptotically negligible compared to $\sum_{i=1}^n E(N_i^o)$. By the central limit theorem,[41] the distribution of N^o tends to be approximately normally distributed with the mean and variance in (4).

It is easy to see that $c^2(N^o)$ can be very large if there are relatively few nodes all in light traffic. If there is a single bottleneck node in

heavy traffic, then $c^2(N^o) \approx 1$, which may also be regarded as significantly different from zero. This indicates that the FPM method might not be desirable with a bottleneck node (see Section IX). However, if there is no bottleneck node and if there are several nodes with at least moderate traffic intensity (and any number of other nodes), then $c^2(N^o)$ will not be large. For example, if there are six nodes with $\rho_i = 2/3$ for each i, then $c^2(N^o) = 0.25$.

2.2 Several servers

It is also worth noting that the equilibrium distribution at a node usually tends to become less variable as the number of servers increases, so that the single-server case we have just considered tends to be the worst case for variability. This is perhaps not true for IS nodes, which often are "delay" nodes. If λ_i is the arrival rate, μ_i is the individual service rate, and $\alpha_i = \lambda_i/\mu_i$ at node i, then since IS nodes have a Poisson distribution,

$$E(N_i^o) = \alpha_i, \qquad \text{Var}(N_i^o) = \alpha_i \quad \text{and} \quad c^2(N_i^o) = 1/\alpha_i.$$

If λ_i and μ_i are the same as in the single-server case, then so is $c^2(N_i^o)$. If λ_i/μ_i tends to increase as the number of servers increases, then $c^2(N_i^o)$ decreases as well. In an M/M/s queue, it is possible to show that $c^2(N_i^o)$ decreases as s increases and converges to 0 as $s \to \infty$, provided that we fix the probability of delay (see Ref. 42). With $\rho_i = \lambda_i/\mu_i s_i$, this is tantamount to having $(1 - \rho_i)\sqrt{s_i} \to \beta_i$, $0 < \beta_i < 1$, as s_i increases. In other words, if we only increase s_i, then ρ_i decreases and λ_i/μ_i remains unchanged, but if we adjust ρ_i as we change s_i to reflect the corresponding congestion, then the distribution tends to concentrate. In particular, we then have $E(N_i^o) \to \infty$, $\text{Var}(N_i^o) \to \infty$, and $c^2(N_i^o) \to 0$ as $s \to \infty$.

2.3 Practical implications

The rather informal analysis in this section indicates that the population variability measured by $c^2(N^o)$ in an open network will often be surprisingly small if (1) the network has quite a few nodes, (2) the network is not in light traffic, and (3) the network is not dominated by one or two bottleneck nodes. Perhaps the most important idea is the possibility of using $c^2(N^o)$ to help determine whether an open model or a closed model is more appropriate. In an application it seems appropriate to measure the real system and estimate the population mean and variance. Then estimate $c^2(N^o)$ for the open model to judge the quality of the fit.

2.4 Reducing the variability of the arrival processes

As mentioned in Section I, we might try to improve the open-model

approximation of a closed model by artificially reducing the variability of the arrival processes in the open model. For example, we might replace each M/M/1 queue by an $E_k/M/1$ queue, where E_k is an Erlang distribution with the same mean. Of particular interest is the limiting case as $k \to \infty$, D/M/1. It is significant that indeed both $\mathrm{Var}(N_i^o)$ and $c^2(N_i^o)$ decrease as we increase k; in fact, $\mathrm{Var}(N_i^o)$ and $c^2(N_i^o)$ for D/M/1 are the least possible values among all GI/M/1 queues with the same arrival rate and service rate. To see this, recall that for a GI/M/1 queue

$$\mathrm{Var}(N_i^o) = \rho_i(1 - \rho_i + \sigma_i)/(1 - \sigma_i)^2 \quad \text{and}$$
$$c^2(N_i^o) = (1 - \rho_i + \sigma_i)/\rho_i, \tag{5}$$

where σ_i is the probability of delay, which is the root of the equation

$$\phi_i(\mu_i(1 - \sigma_i)) = \sigma_i \tag{6}$$

for

$$\phi_i(s) = \int_0^\infty e^{-st} dF_i(t) \tag{7}$$

with F_i the interarrival-time cdf of the GI/M/1 model for node i (see II.3 of Ref. 43).

The relationship is appropriately expressed in terms of stochastic orderings. We say that one random variable X_1 is less than or equal to another X_2 in the sense of *stochastic order* (denoted by $X_1 \leq_{st} X_2$), *increasing convex order* (denoted by $X_1 \leq_{ic} X_2$), and *convex order* (denoted by $X_1 \leq_c X_2$), respectively, if $Eg(X_1) \leq Eg(X_2)$ for all nondecreasing, nondecreasing convex, and convex real-valued functions g for which the expectations are well defined (see Sections 1.3 through 1.5 of Ref. 24). Since $g(x) = x$ and $g(x) = -x$ are both convex, convex order implies equal means. With equal means, convex order is equivalent to increasing convex order. It is significant that $D \leq_c E_{k+1} \leq_c E_k \leq_c M \leq_c H_2$ for random variables with a common mean. (H_2 is a hyperexponential distribution, the mixture of two exponential distributions.)

Let W be the equilibrium waiting time before beginning service. Stoyan and Stoyan showed that $W_1 \leq_{ic} W_2$ in two GI/G/1 queues with common service-time distribution when $X_1 \leq_c X_2$, where X_i is the generic interarrival time in system i.[44] For the special case of GI/M/1, Rolski and Stoyan showed that $W_1 \leq_{st} W_2$ under the same condition.[45] Since $\sigma_i = P(W_i > 0)$, $\sigma_1 \leq \sigma_2$ and, by (5), $\mathrm{Var}(N_1) \leq \mathrm{Var}(N_2)$ and $c^2(N_1) \leq c^2(N_2)$. Since $EX \leq_c X$ for any X, these quantities are minimized among GI/M/1 queues by the D/M/1 case.

Table II shows how the variability is reduced by comparing the

Table II—A comparison of congestion measures for the M/M/1 and D/M/1 queues[67]

Traffic Intensity, ρ_i	M/M/1		D/M/1		
	EN_i^o	$c^2(N_i^o)$	Delay Probability, σ_i	EN_i^o	$c^2(N_i^o)$
0.10	0.11	10.00	0.000	0.10	9.00
0.20	0.25	5.00	0.007	0.20	4.04
0.30	0.43	3.33	0.041	0.31	2.47
0.40	0.67	2.50	0.107	0.45	1.77
0.50	1.00	2.00	0.203	0.63	1.41
0.60	1.50	1.67	0.324	0.89	1.21
0.70	2.33	1.43	0.467	1.31	1.10
0.80	4.00	1.25	0.629	2.16	1.04
0.90	9.00	1.11	0.807	4.66	1.01
0.95	19.00	1.05	0.902	9.69	1.002
0.98	49.90	1.02	0.960	24.50	1.000

principal congestion measures for the M/M/1 and D/M/1 queues. The variability is reduced the most for traffic intensities near 0.5; e.g., for $\rho_i = 0.5$ the D/M/1 value is 70 percent of the M/M/1 value. [From heavy traffic theory,[42] we know that, as $\rho_i \to 1$, $c^2(N_i^o)$ approaches 1 for all GI/M/1 queues and, for D/M/1, EN_i^o approaches one-half the M/M/1 value.] This analysis shows that the proposed technique for refining the approximations by artificially reducing the variability parameters of the arrival processes would indeed reduce the variability of the number of jobs at each node. It also indicates by how much. Since the means would also decrease, this device would also increase the throughput with the FPM method. However, it remains to determine how much to reduce the variability of arrival processes and whether this will produce a good general approximation procedure for network models.

III. COMPARING THROUGHPUTS IN BALANCED NETWORKS

3.1 The closed model with single-server nodes

In this section, we compare the throughput in a closed model with the throughput in the associated open model using the FPM method. We still consider the Markov Jackson models, but for simplicity we restrict attention to single-server nodes. Given a closed model containing n single-server nodes and a fixed population, K, construct the associated open model by removing the arrival process to one node, say the first node, and replacing it by an external Poisson arrival process with sufficiently low arrival rate to have stability. Let the original arrivals to this entry node in the closed model leave the system. Then solve the traffic equations to obtain the arrival rates λ_i and traffic intensities ρ_i for each node i. Note that $n - 1$ of the n

traffic-rate equations for the original closed model are the same as for the new open model.[6] Since there is one degree of freedom in the traffic-rate equations for the closed model, the arrival rates λ_i calculated for the open model are legitimate relative traffic rates for the closed model; i.e., the ratio of the arrival rates at any two nodes thus is identical in the closed and open models.

Let N_i^o and N^o be the equilibrium number of jobs at node i and in the entire system for the open model, and let N_i^c and N^c be the corresponding quantities for the closed model. Obviously, $N^c = K$. For the open network with the given external arrival rate, the expected equilibrium number of jobs is

$$E(N^o) = \sum_{i=1}^{n} (\rho_i/(1 - \rho_i)). \tag{8}$$

It is somewhat remarkable that the equilibrium distribution of the number of jobs at each node in the original closed model can be found by considering the associated open model we have constructed, even though the arrival process we have removed is not a Poisson process. The equilibrium distribution of the numbers of jobs at each node in the original closed model can be expressed exactly in terms of the solution for the open model constructed above;[1-6] it is

$$P(N_i^c = k_i,\ 1 \le i \le n) = \frac{P(N_i^o = k_i,\ 1 \le i \le n)}{P(N^o = K)} = G \prod_{i=1}^{n} \rho_i^{k_i}, \tag{9}$$

where G is the normalization constant or partition function chosen so that the probabilities sum to one over the set of n-tuples $(k_1, k_2, \cdots, k_n)$ such that $k_1 + k_2 + \cdots + k_n = K$. [Of course, this is partly explained by the fact that ratio of arrival rates at any two nodes are the same in the open and closed models. The one remaining degree of freedom, the arbitrary arrival rate in the open model, cancels in the division in (9).]

The associated throughput in the closed model, say θ^c, is then the flow through the designated node, i.e., the utilization u_1^c times the service rate μ_1:

$$\theta^c = u_1^c \mu_1 = P(N_1^c > 0)\mu_1. \tag{10}$$

As usual, the throughput can be obtained by calculating the normalization constant recursively over smaller populations and subsets of nodes (see Section 5.5 of Ref. 6).

3.2 The special case of a balanced network

From (8) through (10), it is clear that the relation between the throughput and the total population is much more elementary for open models; for open models, we only need to know the expected total population, not the detailed distribution at the nodes. We make

an interesting explicit comparison, by considering the special case in which the traffic intensities at all the nodes are identical, say ρ. For the open model, (8) reduces to $EN^o = n\rho/(1 - \rho)$, so that if we set $EN^o = K$, we obtain $\rho = K/(K + n)$. Thus, the utilization u_1^o and throughput θ^o in the open model are

$$u_1^o = K/(K + n) \quad \text{and} \quad \theta^o = \lambda_1 = K\mu_1/(K + n) \tag{11}$$

because all external arrivals but no internal arrivals go to the designated first node.

On the other hand, for the closed model,

$$1/G\rho^K = A_{K,n} = \binom{n + K - 1}{K}, \tag{12}$$

the number of ways K indistinguishable objects can be placed into n cells (p. 38 of Ref. 41). Similarly,

$$P(N_1^c = 0) = A_{K,n-1}/A_{K,n} = (n - 1)/(n + K - 1), \tag{13}$$

so that the utilization and throughput in the closed network are

$$u_1^c = K/(K + n - 1) \quad \text{and} \quad \theta^c = K\mu_1/(n + K - 1). \tag{14}$$

From (11) and (14), we see that the two throughputs are very similar. Moreover, $\theta^c > \theta^o$, so that θ^o is a conservative estimate of θ^c. (This is always true; see Section IV.)

It is significant that a good approximation for the throughput θ^c in the closed model immediately provides a good approximation for the utilizations of all the nodes. As we have indicated above, the closed and open models are linked together in a very important way: The ratio of any two utilizations is always identical in both models; i.e.,

$$u_i^o/u_j^o = u_i^c/u_j^c \tag{15}$$

for all i and j.

The difference between the two utilizations, say Δ, which for the balanced model is the difference between the throughputs in (11) and (14) normalized by dividing by the service rate μ_1, is

$$\Delta \equiv u_1^c - u_1^o = (\theta^c - \theta^o)/\mu_1 = K/(n + K)(n + K - 1). \tag{16}$$

Note that $\theta^c = \mu_1$, $\theta^o = \mu_1/2$, and $\Delta = 1/2$ when $K = n = 1$. We conjecture that $\Delta \geq 1/2$, in general. The difference Δ is small if either n or K (especially n) is large but not if n and K are both small. Representative values of n, K, θ^c, θ^o, and Δ are given in Table III. In Table III the service rate at the entry node is $\mu_1 = 1$. We also describe the population variability in the open model using (4) and (11), from

Table III—A comparison of throughputs for the network of single-server nodes with common relative traffic intensities considered in Section III

Network Parameters		Throughput Measures		Difference in Throughputs	Population Variability
Nodes n	Jobs K	Closed θ^c	Open θ^o	Δ (16)	$c^2(N^o)$
2	2	0.67	0.50	0.17	1.00
2	5	0.83	0.71	0.12	0.70
2	20	0.95	0.91	0.04	0.55
5	2	0.33	0.29	0.04	0.70
5	5	0.56	0.50	0.06	0.40
5	20	0.83	0.80	0.03	0.25
20	2	0.10	0.09	0.01	0.55
20	5	0.21	0.20	0.01	0.25
20	20	0.51	0.50	0.01	0.10

which $c^2(N) = (K + n)/Kn$. The difference between the two models as described by both Δ and $c^2(N^o)$ decreases as n and K increase. Table III quantifies the differences.

Formulas (11) through (16) indicate that the FPM approximation for the throughput will be good if either the population, K, or the number of nodes, n, is large, but with single-server nodes it seems much better to have n large. The quality of the approximate queue-length distributions computed by the FPM method often deteriorates when there are nodes with high utilizations and few servers. Example 1 in Section I is ideal for the FPM method; both K and n are large ($K = n = 20$), but the utilizations are not. The finite-waiting-room refinement in Section 1.3 is useful for the small models.

3.3 Simple approximations for unbalanced networks

We can use the results for balanced networks to obtain simple approximations for unbalanced networks. A simple rough approximation, say θ^c_{approx}, for the throughput in a closed network with K jobs and n single-server nodes with unequal (but not too different) utilizations based on (11) and (14) is

$$\theta^c_{approx} = \theta^o(n + K)/(n + K - 1), \tag{17}$$

where θ^o is obtained from the associated open model using (8), e.g., by simple search. We would not expect (17) to be good if there is a severe bottleneck node; we would be in serious trouble if we had six nodes, five having relative utilization 1 and the other having relative utilization 3. We also would not count nodes in relatively light traffic; if we had nine nodes, three with relative utilizations 1 and 6 with relative utilization 3, then it would be better to use $n = 6$ in (17).

IV. SUPPORTING THEORY FOR COMPARING THE MODELS

4.1 A lower bound for the closed model

For the special example in Section 3.2, we saw from (11) and (14) that $\theta^c > \theta^o$. In general, we should expect the throughput to be greater in the closed model because it is intuitively obvious that N_i^c is less variable than N_i^o. Given the same mean, N_i^o is evidently more likely to assume both very large values and very small values, so that we should have $P(N_i^o = 0) > P(N_i^c = 0)$. Of course, we need not actually have $EN_i^o = EN_i^c$ when $EN^o = N^c$, but this is the idea.

In this section we justify this reasoning. We assume that the open model is constructed from the closed model as described in Section 3.1. We consider the Markov Jackson network with one job class and multiserver nodes as specified in Section I, but it is significant that the throughput comparisons extend to Markov networks with multiple job classes and more general state-dependent service rates at the nodes. Some of these comparisons also apply to the finite-waiting-room approximation introduced in Section 1.3. To avoid complicated notation, we only discuss these extensions in remarks after Theorem 4.

Theorem 1: If $EN^o \le N^c$, then $\theta^o \le \theta^c$ and $u_i^o \le u_i^c$ for all i.

For the special case in which all nodes are either single-server or IS nodes, Zahorjan[22] has given a nice proof of Theorem 1. We give a different argument that allows us to treat more general nodes, e.g., multiserver nodes, and obtain some interesting additional results along the way. To establish Theorem 1, we use notions of *concave ordering*, which are closely related to the convex orderings introduced in Section II (see Section 1.4 of Ref 24). One random variable X_1 is less than or equal to another X_2 in concave (increasing concave) ordering, denoted by $X_1 \le_{cv} X_2$ ($X_1 \le_{icv} X_2$), if $Eg(X_1) \le Eg(X_2)$ for all concave (increasing concave) real-valued functions g for which the expectations are defined. The connection to the convex orderings is simple: $X_1 \le_{cv} X_2$ if and only if $X_1 \ge_c X_2$; $X_1 \le_{icv} X_2$ if and only if $-X_1 \ge_{ic} -X_2$. The following basic characterization for random variables with values in the nonnegative integers is useful: $X_1 \le_{icv} X_2$ if and only if

$$\sum_{k=0}^{n} P(X_1 \le k) \ge \sum_{k=0}^{n} P(X_2 \le k) \tag{20}$$

for all n (see Sections 1.3 through 1.5 of Ref. 24).

As a basis for Theorem 1, we establish the following result.

Theorem 2: If $EN_i^o \le EN_i^c$ for any node i, then $N_i^o \le_{icv} N_i^c$.

In fact, Theorem 2 directly implies Theorem 1 given (15). Let u_i^c and u_i^o be the utilizations of node i in the closed and open models, respectively. They are both defined as the expected number of busy servers; e.g., for the open model,

We can also directly approximate θ^o by replacing the traffic intensities at each node with the average traffic intensity over all nodes. This yields $\theta^o_{approx} = K\mu_1/(K + n)$ as in (11). Due to the convexity of EN_i^o, $\theta^o_{approx} \geq \theta^o$, which is a modification in the correct direction if we wish to approximate θ^c.

Finally, we could combine these two approximations to obtain (14) for unbalanced closed networks, but it appears that this would tend to overestimate θ^c. In fact, (14) has already been shown to be an upper bound for the closed model.[9-11] The simple approximation (17) worked very nicely in Example 1 in Section 1.2.

3.4 Reducing the variability of the arrival processes

As in Section 2.4, we can consider approximations for the closed model obtained by reducing the variability in the open model. Since $EN_i^o = \rho_i/(1 - \sigma_i)$ in the GI/M/1 model, EN_i^o also decreases as σ_i decreases, so reducing the variability of the arrival process at each node increases the throughput θ^o in the open model. (Typical values of EN_i^o for the D/M/1 queue are given in Table II.) If there are n identical GI/M/1 nodes, then instead of (11) we have

$$\theta^o = K\mu_1/(n + xK), \tag{18}$$

where $x = \sigma/\rho$. Since σ depends on ρ via (6), (18) is harder to solve. Moreover, a direct application of the D/M/1 model need not yield good results because (18) can be much greater than (11). For example, if $K = 10$, $n = 16$, and $\mu_1 = 1$, then $\theta^c = 0.40$, while $\theta^o = 0.385$ and 0.50 via (11) and (18), respectively. It remains to determine how to exploit this approach.

3.5 Several servers

It is also interesting to consider networks of n identical multiserver nodes (back with the Markov models). When there are s servers with $1 < s < \infty$, the formulas are rather complicated, but the situation simplifies greatly for $s = \infty$. Then $EN_i^c = EN_i^o = K/n$ and $\theta^c = \theta^o = K\mu_1/n$, so that there is no difference at all. We conjecture that $(\theta^c - \theta^o)/\mu$, decreases in s, which would mean that the single-server case we have examined gives the worst approximation.

For the open model in which each node has s servers and a common traffic intensity, $\rho = \lambda_1/s\mu_1$, θ^o can be approximated by solving

$$n[\rho s + \delta\rho/(1 - \rho)] = K, \tag{19}$$

where δ is the probability of delay at node 1 (which also depends on λ_1) (see Ref. 42). A possible procedure is to approximate δ first and then solve the resulting quadratic equation for λ. One could then iterate, recalculating δ, etc.

IV. SUPPORTING THEORY FOR COMPARING THE MODELS

4.1 A lower bound for the closed model

For the special example in Section 3.2, we saw from (11) and (14) that $\theta^c > \theta^o$. In general, we should expect the throughput to be greater in the closed model because it is intuitively obvious that N_i^c is less variable than N_i^o. Given the same mean, N_i^o is evidently more likely to assume both very large values and very small values, so that we should have $P(N_i^o = 0) > P(N_i^c = 0)$. Of course, we need not actually have $EN_i^o = EN_i^c$ when $EN^o = N^c$, but this is the idea.

In this section we justify this reasoning. We assume that the open model is constructed from the closed model as described in Section 3.1. We consider the Markov Jackson network with one job class and multiserver nodes as specified in Section I, but it is significant that the throughput comparisons extend to Markov networks with multiple job classes and more general state-dependent service rates at the nodes. Some of these comparisons also apply to the finite-waiting-room approximation introduced in Section 1.3. To avoid complicated notation, we only discuss these extensions in remarks after Theorem 4.

Theorem 1: If $EN^o \le N^c$, then $\theta^o \le \theta^c$ and $u_i^o \le u_i^c$ for all i.

For the special case in which all nodes are either single-server or IS nodes, Zahorjan[22] has given a nice proof of Theorem 1. We give a different argument that allows us to treat more general nodes, e.g., multiserver nodes, and obtain some interesting additional results along the way. To establish Theorem 1, we use notions of *concave ordering*, which are closely related to the convex orderings introduced in Section II (see Section 1.4 of Ref 24). One random variable X_1 is less than or equal to another X_2 in concave (increasing concave) ordering, denoted by $X_1 \le_{cv} X_2$ ($X_1 \le_{icv} X_2$), if $Eg(X_1) \le Eg(X_2)$ for all concave (increasing concave) real-valued functions g for which the expectations are defined. The connection to the convex orderings is simple: $X_1 \le_{cv} X_2$ if and only if $X_1 \ge_c X_2$; $X_1 \le_{icv} X_2$ if and only if $-X_1 \ge_{ic} -X_2$. The following basic characterization for random variables with values in the nonnegative integers is useful: $X_1 \le_{icv} X_2$ if and only if

$$\sum_{k=0}^{n} P(X_1 \le k) \ge \sum_{k=0}^{n} P(X_2 \le k) \tag{20}$$

for all n (see Sections 1.3 through 1.5 of Ref. 24).

As a basis for Theorem 1, we establish the following result.

Theorem 2: If $EN_i^o \le EN_i^c$ for any node i, then $N_i^o \le_{icv} N_i^c$.

In fact, Theorem 2 directly implies Theorem 1 given (15). Let u_i^c and u_i^o be the utilizations of node i in the closed and open models, respectively. They are both defined as the expected number of busy servers; e.g., for the open model,

$$u_i^o = E(\min\{N_i^o, s_i\}) = \rho_i s_i = \lambda_i/\mu_i, \tag{21}$$

where, of course, λ_i is the net arrival rate determined by the traffic rate equations plus the external arrival rate, which is, in turn, determined by the FPM requirement that $EN^o = N^c$. Formula (21) is also valid for the closed model.

To prove Theorem 1, we use the following consequence of Theorem 2.

Corollary to Theorem 2: If $EN_i^o \leq EN_i^c$, then $u_i^o \leq u_i^c$.

Proof: Apply Theorem 2 observing that the function in (21) is increasing and concave. □

Proof of Theorem 1: If $EN^o \leq N^c$, then $EN_i^o \leq EN_i^c$ for some i because $\sum_{i=1}^n EN_i^c = EN^c = N^c$. For one such i, $u_i^o \leq u_i^c$ by Theorem 2 and its corollary. By (15), $u_i^o \leq u_i^c$ for all i. Since $\theta^c = u_1^c \mu_1$ and $\theta^o = u_1^o \mu_1$, $\theta^c > \theta^o$ too. □

To prove Theorem 2, we use notions of log-concavity (see p. 70 of Ref. 23). A probability mass function $\{p_k, k \geq 0\}$ is *log-concave* if

$$p_k^2 \geq p_{k+1} p_{k-1}, \qquad k \geq 1. \tag{22}$$

A log-concave distribution is unimodal; moreover, it is strongly unimodal, i.e., the convolution with any unimodal distribution is also unimodal. In fact, for discrete distributions log-concavity, strong unimodality, and the PF_2 (Polya frequency function) property are all equivalent.[23] The equilibrium distribution of any birth-and-death process is log-concave if the birth rates are nonincreasing and the death rates are nondecreasing (see example 5.7F in Ref. 23). Moreover, log-concavity is preserved under convolution. Hence, for each i and m the distributions of N_i^o and $N_1^o + \cdots + N_m^o$ are log-concave. (By example 5.7F in Ref. 23 referred to above, it suffices for the service rate at each node to be a nondecreasing function of the number of jobs present.) It turns out that this is also true for the more complicated distributions in the closed network.

Theorem 3: Let the service rate at each node be a nondecreasing function of the number of jobs present. For any m the distribution of $N_1^c + \cdots + N_m^c$ is log-concave.

Proof: Consider $m = 1$. Since log-concavity is preserved under convolution,[23] the distribution of $N_2^o + \cdots + N_n^o$ is log-concave. Then note that

$$\frac{P(N_1^c = k+1)}{P(N_1^c = k)} = \frac{P(N_1^o = k+1)P(N_2^o + \cdots + N_n^o = K-k-1)}{P(N_1^o = k)P(N_2^o + \cdots + N_n^o = K-k)}, \tag{23}$$

with the right-hand side being the product of two ratios, both decreasing in k. A similar argument applies to $m > 1$. □

From (23), we see that in some sense the distribution of N_i^c is more log-concave than the distribution of N_i^o. We now formalize this notion.

Definition 1: One probability mass function $\{p_k^1, k \geq 0\}$ is said to be log-concave *relative* to another $\{p_k^2, k \geq 0\}$ if $(p_{k+1}^1 p_k^2)/(p_k^1 p_{k+1}^2)$ is nonincreasing in k.

From (23) it is obvious that N_i^c is log-concave relative to N_i^o. We now show that this supplies what we need for Theorem 2. The key property is that relative log-concavity implies that the ratio p_k^1/p_k^2 is unimodal.[23]

Theorem 4: If the distribution of a random variable X_2 is log-concave relative to the distribution of another random variable X_1 and $EX_1 \leq EX_2$, then $X_1 \leq_{\text{icv}} X_2$.

Proof: Our goal is to verify (20). We first show that $P(X_1 = 0) \geq P(X_2 = 0)$. If not, then the relative log-concavity implies that there is a k_0 such that $P(X_1 = k) < P(X_2 = k)$ for all $k \leq k_0$ and no $k > k_0$. This would make X_1 stochastically larger than X_2, implying that $EX_1 > EX_2$, which contradicts an assumption. Hence, $P(X_1 = 0) \geq P(X_2 = 0)$. Next let k_1 be the first k, if any, such that $P(X_1 \leq k_1) \leq P(X_2 \leq k_1)$. By the relative log-concavity, we must have $P(X_1 \leq k) \leq P(X_2 \leq k)$ for all $k \geq k_0$. Since $EX_i = \sum_{k=0}^{\infty} P(X_i \geq k)$,

$$EX_2 - EX_1 = \sum_{k=0}^{\infty} [P(X_1 \leq k) - P(X_2 \leq k)] > 0$$

and

$$\sum_{k=0}^{n} P(X_1 \leq k) \geq \sum_{k=0}^{n} P(X_2 \leq k), \qquad n \geq 0,$$

which establishes (20). □

Proof of Theorem 2: By (23), the distribution of N_i^c is log-concave relative to the distribution of N_i^o according to Definition 1. By Theorem 4, $N_i^o \leq_{\text{icv}} N_i^c$. □

Remarks: 1. In a network made up entirely of infinite-server nodes, we have $u_i^o = u_i^c$ for all i and $\theta^o = \theta^c$, so that we cannot have strict inequality in Theorems 1 and 2.

2. Theorems 2 through 4 apply to the FPM/FWR method introduced in Section 1.3. Let $\bar{N}_i^o$ be the equilibrium number at node i by this method. It is easy to see that N_i^c is log-concave relative to $\bar{N}_i^o$, which in turn is log-concave relative to N_i^o. Hence, if $EN_i^o \leq E\bar{N}_i^o$, then $N_i^o \leq_{\text{icv}} \bar{N}_i^o$; if $E\bar{N}_i^o \leq EN_i^c$, then $\bar{N}_i^o \leq_{\text{icv}} N_i^c$. However, this does not yield a proof of the analog of Theorem 1 because the relationship (15) is lost. We do obtain the lower bound (3), though.

3. As indicated at the beginning of this section, Theorems 1 through 4 extend to multiple job classes. There are many different ways to define the class structure, but we shall use only basic properties that have been established for the Markov models.[5,6] For the open model, the vector of jobs at the nodes without identifying the classes has the same equilibrium distribution as when there is only a single class, and given any number of jobs at node i in equilibrium, each job is of class j with some probability p_{ij}, independently of the other jobs. In other words, if N_i^o is the total number of jobs at node i and N_{ij}^o is the number of class j jobs at node i, then N_{ij}^o is obtained from N_i^o Bernoulli trials with probability p_{ij}:

$$P(N_{ij}^o = k) = \sum_{n=k}^{\infty} P(N_i^o = n) \binom{n}{k} p_{ij}^k (1 - p_{ij})^{n-k}. \tag{24}$$

The key property is that the distribution of N_{ij}^o in (24) is log-concave whenever the distribution of N_i^o is log-concave. This result is intuitively reasonable, but not so easy to prove. The result is established in Theorem 2 of Ref. 46. Given that N_{ij}^o has a log-concave distribution and that $N_{i_1 j}^o$ is independent of $N_{i_2 j}^o$ when $i_1 \neq i_2$, it is easy to extend all of the previous results in this section to multiple job classes. For example, the extension of Theorem 1 states that the utilizations of each class at each node are ordered, i.e., $u_{ij}^o \leq u_{ij}^c$ for all i and j, if the expected class populations are ordered, i.e., $\sum_{i=1}^n EN_{ij}^o \leq \sum_{i=1}^n N_{ij}^c = K_j$ for all j.

4. We have indicated that p^1 being log-concave relative to p^2 implies that p_k^1/p_k^2 is unimodal in k. We call this relationship Uniform Conditional Variability Order (UCVO), provided that p^1 and p^2 are not stochastically ordered, because all conditional distributions, conditioning on a common subset, are either ordered again by UCVO or are ordered by ordinary stochastic order. This property parallels uniform conditional stochastic order,[47,48] and is studied further elsewhere.[49]

4.2 Dependence in the closed model

So far in this section (in Theorem 2 and Remark 4 above), we have shown how to express the idea that the distribution of the number of jobs at each node is less variable in the closed model than in the open model, but we have yet to describe the joint distribution at several nodes. Unlike in the open model, where the marginal distributions are independent, in the closed model the marginal distributions are dependent. If there are more jobs in one subset of nodes, then there should be fewer jobs at another disjoint subset of nodes. The population constraint obviously should make the populations at different nodes negatively correlated. We can make these ideas precise using recently developed concepts of negative dependence.

One concept of negative dependence is the Multivariate Reverse

Rule distribution (MRR_2), which was introduced by Karlin and Rinott.[25] Let p be a multivariate probability mass function on the n-fold product of the nonnegative integers. The distribution p is said to be MRR_2 if

$$p(x \vee y)p(x \wedge y) \leq p(x)p(y) \tag{25}$$

for all $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ in $\{0, 1, \cdots\}^n$, where

$$x \vee y = (\max\{x_1, y_1\}, \cdots, \max\{x_n, y_n\}) \quad \text{and}$$

$$x \wedge y = (\min\{x_1, y_1\}, \cdots, \min\{x_n, y_n\}).$$

In contrast, if p satisfies (25) with the inequality reversed, p is said to be Multivariate Totally Positive (MTP_2).[50] In both cases it suffices to check (25) for x and y differing in only two components.

Unlike MTP_2 distributions,[50] the marginal distributions of an MRR_2 distribution need not be MRR_2. Moreover, even having the marginal distributions all MRR_2 is not strong enough to deduce some of the desired multivariate inequalities. Karlin and Rinott[25] proposed one way to cope with this difficulty, by introducing a special subclass called the strongly MRR_2 ($SMRR_2$) distributions. An *n-dimensional* probability mass function p is $SMRR_2$ if the $(n - m)$-dimensional function $\sum p(x_1, \cdots, x_n)\, \phi_1(x_{j_1}) \cdots \phi_m(x_{j_m})$ is MRR_2 for all m and all m-tuples of indices $(j_1, \cdots, j_m)$, where the sum is over all $(x_{j1}, \cdots, x_{jm})$ and ϕ_i is log-concave (PF_2) for each i.

Block, Savits, and Shaked[26] introduced a convenient structural condition (condition N) that implies $SMRR_2$. A random vector $(X_1, \cdots, X_n)$ satisfies condition N if there is a vector of $n + 1$ independent random variables $(Y_0, Y_1, \cdots, Y_n)$ each with a PF_2 density (or mass function) such that $(X_1, \cdots, X_n)$ is distributed the same as $[(Y_1, \cdots, Y_n) \mid Y_0 + \cdots + Y_n = s]$ for some s. It is easy to see that condition N applies to the closed models as a special case; just set $Y_0 = 0$ and $s = K$.

Another concept of negative dependence was proposed by Joag-Dev and Proschan.[27] They call random variables $X_1, \cdots, X_n$ and their joint distribution Negatively Associated (NA) if, for every pair of disjoint subsets A_1 and A_2 of the index set $\{1, 2, \cdots, n\}$, the covariance

$$\operatorname{cov}(f((X_i, i \epsilon A_1)), g((X_j, j \epsilon A_2))) \leq 0 \tag{26}$$

for all nondecreasing real-valued functions f and g defined on R^{k_1} and R^{k_2}, where k_i is the cardinality of A_i. Their Theorem 8 directly implies that $(N_1^c, \cdots, N_n^c)$ is negatively associated. We collect these properties in Theorem 5.

Theorem 5: The vector $(N_1^c, \cdots, N_n^c)$ is negatively associated, satisfies condition N, is $SMRR_2$, and has all marginal distributions MRR_2.

Proof: Since $(N_1^c, \cdots, N_n^c)$ is distributed as $(N_1^o, \cdots, N_n^o | N_1^o + \cdots + N_n^o = K)$, condition N in Ref. 26 and the sufficient condition for the NA property in Theorem 2.8 of Ref. 27 are immediate. As noted in the Remark at the end of Section IV of Ref. 26, condition N implies $SMRR_2$, which in turn implies that all marginals are MRR_2. □

Remark: It is elementary to directly verify that all marginals are MRR_2.

Many important consequences of Theorem 5 are described in Refs. 25 through 27. We give some illustrative examples.

Corollary 1 to Theorem 5: Suppose that ϕ_i are all nondecreasing or all nonincreasing functions on the nonnegative integers. Then, for any k, $1 \le k \le n$,

$$E\{\phi_1(N_1^c)\phi_2(N_2^c) \cdots \phi_n(N_n^c)\}$$
$$\le E\{\phi_1(N_1^c) \cdots \phi_k(N_k^c)\} \times E\{\phi_{k+1}(N_{k+1}^c) \cdots \phi_n(N_n^c)\}.$$

Proof: Apply (1.5) of Ref. 25, noting that the PF_2 property is not used there.

Corollary 2 to Theorem 5: For any $m \le n$ and any m-tuple $(k_1, \cdots, k_m)$,

$$P(N_i^c \le k_i, 1 \le i \le m) \le \prod_{i=1}^{m} P(N_i^c \le k_i)$$

and

$$P(N_i^c \ge k_i, 1 \le i \le m) \le \prod_{i=1}^{m} P(N_i^c \ge k_i).$$

Proof: Apply Corollary 1. □

Remark: Theorem 5 and its corollaries describe multivariate dependence for a single job class. The multiple-job-class closed Markov network suggests a natural generalization of condition N in Ref. 26, which we call condition QN. A random vector $X \equiv (X_{ij}: 1 \le i \le n, 1 \le j \le m)$ in R^{mn} satisfies condition QN if it is appropriately related to another random vector. The other random vector is $Y \equiv (Y_{ij}: 0 \le i \le n, 1 \le j \le m)$ in $R^{m(n+1)}$ such that (1) the subvectors $(Y_{0j}: 1 \le j \le m)$, $(Y_{1j}: 1 \le j \le m)$, $\cdots$, $(Y_{nj}: 1 \le j \le m)$ are mutually independent, (2) the random variables $\sum_{j=1}^{m} Y_{ij}$ have a PF_2 density or mass function for each i, and (3) given $\sum_{j=1}^{m} Y_{ij}$, $(Y_{i1}, \cdots, Y_{im})$ has a multinomial distribution for each i. We say that X satisfies condition QN if X is distributed as $(Y_{ij}: 1 \le i \le n, 1 \le j \le m)$ conditional on $\sum_{i=0}^{n} Y_{ij} = s_j$, $1 \le j \le m$, for some m-tuple $(s_1, \cdots, s_m)$. For $m = 1$, of course condition QN reduces to condition N. Condition QN is being investigated; a discussion of its properties is intended for a future paper.

With closely related stochastic comparison concepts, we can also describe what happens when we increase the population in a closed network. Let $(N_1^c(K), \cdots, N_n^c(K))$ be the equilibrium vector of jobs at each node as a function of the total population K. Naturally, we expect it to be increasing in K in some sense. In fact, this is true in a very strong sense. Following Karlin and Rinott,[50] we say that one multivariate probability mass function p_1 is less than or equal to another p_2 in the sense of multivariate Monotone Likelihood Ratio (MLR), and we write $p_1 \leq_{lr} p_2$, if

$$p_1(x)p_2(y) \leq p_1(x \wedge y)p_2(x \vee y) \tag{27}$$

for all $x = (x_1, \cdots, x_n)$ and $y = (y_1, \cdots, y_n)$ in $\{0, 1, \cdots\}^n$. The *MLR* order is a generalization of MTP_2 because $p \leq_{lr} p$ if and only if p is *MTP*.[50] *MLR* order implies stochastic order for the original distributions (i.e., $p_1 \leq_{st} p_2$) and also for all conditional distributions conditioning on sublattices.[48]

The probability mass functions of the full vector $[N_1^c(K), \cdots, N_n^c(K)]$ are not usefully compared by (27) for different K because they have disjoint support sets. However, we can usefully compare the marginal distributions. There is a further complication, however, because (27) will obviously fail when x and y are in the support of both distributions but $x \vee y$ is not. However, the ordering (27) does hold over every sublattice of the support.

Theorem 6: Given a closed model with n nodes, $[N_1^c(K), \cdots, N_{n-1}^c(K)]$ *is nondecreasing in K in the MLR ordering in the sense that the support set* $\{(k_1, \cdots, k_{n-1}): k_1 + \cdots + k_{n-1} \leq K\}$ *in* $\{0, 1, \cdots\}^{n-1}$ *is nondecreasing in K and* (27) *holds for K and K + 1 as an equality whenever the sum of the components of* $x \vee y$ *is less than or equal to* $K + 1$.

Proof: It suffices to establish (27) for x and y differing by one in just two indices, say 1 and 2. Let $x = (k_1 + 1, k_2, k_3, \cdots, k_{n-1})$ and $y = (k_1, k_2 + 1, k_3, \cdots, k_{n-1})$. Let p_K be the probability mass function of $[N_1^c(K), \cdots, N_n^c(K)]$. Then (27) holds as an equality provided that $k_1 + k_2 + k_{n-1} + 2 \leq K + 1$ because

$$p_K(x)\, p_{k+1}(y)/p_K(x \wedge y)\, p_{K+1}(x \vee y)$$

$$= \frac{P\left(N_n^o = K - \sum_{j=1}^{n-1} k_j - 1\right) P\left(N_n^o = (K+1) - \sum_{j=1}^{n-1} k_j - 1\right)}{P\left(N_n^o = K - \sum_{j=1}^{n-1} k_j\right) P\left(N_n^o = (K+1) - \sum_{j=1}^{n-1} k_j - 2\right)}$$

$$= 1. \quad \square$$

Corollary 1 to Theorem 6: For each i and K, $N_i^c(K) \leq_{lr} N_i^c(K+1)$.

Corollary 2 to Theorem 6: The utilizations $u_i^c(K)$ are increasing in K.

Proof: Apply Corollary 1 with the increasing function in (20). □

Corollary 3 to Theorem 6: The conditional distribution of $[N_1^c(K), \cdots, N_{n-1}^c(K)]$ given any sublattice of $\{0, 1, \cdots, m\}^{n-1}$ with maximal element $(k_1, \cdots, k_{n-1})$ is independent of K for $K \geq \sum_{j=1}^{n-1} k_j$.

Corollary 4 to Theorem 6: $P(N_i^c(K) \geq m \mid a_j \leq N_j^c(K) \leq b_j, 1 \leq j \leq n-1)$ is independent of K for $K \geq \sum_{j=1}^{n-1} b_j$.

Proof: Apply Corollary 3. □

V. LIMITS FOR GROWING NETWORKS

In this section we provide some additional theory to show that the FPM method tends to perform well as an approximation for closed models when the network is large. For particular kinds of growing networks we prove that the FPM method is asymptotically correct as the number of nodes increases. We assume that the population is nK when there are n nodes. As we indicated in Section 1.2, the basic idea here is due to Gordon and Newell (see pp. 261 through 265 of Ref. 2), but we formulate and prove a limit theorem.

To be precise, we must specify how the network topology and other model parameters change as the closed network grows. If the connectivity increases as the closed network grows, so that the departure processes are split into many components and the arrival processes are superpositions of many components, then it is usually possible to show that any fixed finite subset of nodes in the closed network behaves asymptotically as mutually independent queues (with mutually independent Poisson arrival processes). This is an even stronger form of independence than in the open model because it applies to the time-dependent stochastic processes as well as the equilibrium distribution. A simple example is the n-node network with routing of departures from every node to all other nodes with equal probability. Growing networks with increasing connectivity can be treated by classical limit theorems for superposition and thinning.[51,52]

Motivated by Example 1 in Section 1.2, we formulate a limit theorem in which the connectivity does not grow with n. We define our sequence of closed models as follows. We start with a general open Markov product-form network having q nodes, p job classes, and a p-tuple of independent Poisson processes determining the arrivals (one for each class). We then replicate this network n times, letting the departures from network k be the arrivals to network $k+1$. We let the routing probabilities at each subnetwork be identical. Finally, we make it a closed model by replacing the external arrival process to network 1 by the departure process from network n and stipulating that there are nK_j customers of class j, $1 \leq j \leq p$. A symmetric closed cyclic network

is the special case in which the initial building-block network has one node. Example 1 in Section 1.2 can be regarded as the special case in which the initial building-block network is a network with two single-server nodes in series, one with mean service time 1.0 and the other with mean service time 1.2. The numerical calculations were for $n = 10$.

The following result shows that the FPM method is asymptotically correct for such cyclically growing networks as $n \to \infty$. In Remark 1 following the proof, we indicate that Theorem 7 also applies to growing networks with increasing connectivity. Let $N^c_{ijk}(n)$ be the number of class j jobs at node i of the kth subnetwork in the nth closed model; let $N^o_{ijk}(\lambda)$ be the number of class j jobs at node i of the kth subnetwork in the associated open model having independent Poisson arrival processes with rate vector λ at the first subnetwork, with all departures from the nth subnetwork leaving the system.

Theorem 7: For any k_0 and (qpk_0)-tuple $(m_{111}, \cdots, m_{qpk_0})$,

$$\lim_{n\to\infty} P(N^c_{ijk}(n) = m_{ijk}: 1 \le i \le q, 1 \le j \le p, 1 \le k \le k_0) = \prod_{k=1}^{k_0} \prod_{i=1}^{q} P(N^o_{ij1}(\lambda) = m_{ijk}: 1 \le j \le p) \equiv Z,$$

where $\lambda \equiv (\lambda_1, \cdots, \lambda_p)$ is the FPM solution for the original q-node open building-block network.

Proof: Let $M_j = \sum_{k=1}^{k_0} \sum_{i=1}^{q} m_{ijk}$ for $1 \le j \le p$. As in (9),

$$P(N^c_{ijk}(n) = m_{ijk}: 1 \le i \le q, 1 \le j \le p, 1 \le k \le k_0) = \frac{Z \cdot P\left(\sum_{k=k_0+1}^{n} \sum_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j - M_j: 1 \le j \le p\right)}{P\left(\sum_{k=1}^{n} \sum_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j: 1 \le j \le p\right)}$$

for any vector of external arrival rates λ for which there is stability. (Z is defined in the statement of Theorem 7.) For the special case in which λ is the FPM vector, we can apply the local central limit theorem, pp. 75 through 79 of Spitzer,[12] to obtain our desired result. Suppose that λ is such that $E \sum_{i=1}^{q} N^o_{ij1}(\lambda) = K_j$ for each j. Since $\sum_{k=1}^{n} (\sum_{i=1}^{q} N^o_{ijk}(\lambda) - K_j)$ is the jth component of the nth partial sum of i.i.d. random p-tuples where each component has mean 0, there exists α, $0 < \alpha < \infty$, such that

$$\lim_{n\to\infty} n^{p/2} P\left(\sum_{k=1}^{n} \sum_{i=1}^{q} N^o_{ijk}(\lambda) = nK_j + a_j: 1 \le j \le p\right) = \alpha$$

for any p-tuple $(a_1, \cdots, a_p)$. (It is easy to see that the aperiodicity

requirement in the local central limit theorem is satisfied.) We thus obtain the desired result by multiplying both the numerator and denominator by $n^{p/2}$ and letting $n \to \infty$. □

Remarks: 1. Theorem 7 and its proof also apply to other kinds of growing networks in which the connectivity does increase. For example, consider the symmetric n-node network in which each node has the same external Poisson arrival process with rate λ_{oj} and probability r_j of departures leaving the system for class j, $1 \le j \le p$, independent of n. Also, let departures staying within the network be routed to all other nodes with equal probability. Then the arrival rate of class j at each node is $\lambda_{oj}/(1 - r_j)$, independent of n. Then the equilibrium vectors of jobs at any node in the open model are the same for all nodes and are independent of n, so that Theorem 7 and its proof applies with $q = 1$.

A more complex symmetric growing model with increasing connectivity that can be treated the same way is obtained by replacing each node in this example by a q-node subnetwork. The departures from each q-node subnetwork staying within the system would be routed to each possible q-node subnetwork with equal probability. Each q-node subnetwork would also have its own external arrival processes. Then the equilibrium vector of jobs of each class at each node in the open model is the same for all q-node subnetworks and is independent of n.

2. Gordon and Newell propose a refinement to the FPM approximation for large networks, (27) in Ref. 2, which is obtained by approximating the probabilities involving the large partial sums in the numerator and denominator of by the normal density function. This is justified by the Remark on p. 77 of Ref. 12.

VI. THE FPM METHOD WITH A DECOUPLING INFINITE-SERVER NODE

We now consider the special case of a closed network with an IS node. As in Section V, we allow p different job classes. We introduce this added complexity here because our algorithm is particularly well suited to cope with it. Let each class have its own population and routing probabilities. Let there be $q + 1$ nodes with node $q + 1$ being an IS node and assume that it is visited by every class. (It would suffice to have different IS nodes visited by different classes. The other nodes visited by any class might include IS nodes too; the designated IS node has especially low service rates.) Let μ_j be the individual service rate of class j at node $q + 1$. Let K_j be the given fixed population of class j, $1 \le j \le p$. (We are now in the setting of Ref. 14, except that we are allowing multiserver nodes.)

We now modify the original closed model by removing the departure

processes for the p classes from the designated IS node and replacing them by p independent external Poisson arrival processes to the remaining q nodes. Let λ_j be the rate of the Poisson process for class j and let $\underline{\lambda} = (\lambda_1, \cdots, \lambda_p)$.

Let $N_{ij}^o(\underline{\lambda})$ be the equilibrium steady number of customers of class j at node i in the q-node open network without the IS node based on the Poisson external arrival processes with rate vector $\underline{\lambda}$. (We have changed the notation somewhat to emphasize the dependence on $\underline{\lambda}$.) We use the designated IS node to determine the appropriate arrival-rate vector $\underline{\lambda}$. Since the arrival rates equal the departure rates in the open network, the departure rate of class j from the q-node open subnetwork is also λ_j. Since these departures all leave from the designated IS node, we use the IS node to enforce consistency. In particular, we require that the arrival rate be equal to the departure rate for each class at the IS node, i.e.,

$$\lambda_j = \left(K_j - \sum_{i=1}^{q} EN_{ij}^o(\underline{\lambda})\right) \mu_j \tag{28}$$

for each j.

Since the expected equilibrium number of customers in a G/GI/∞ model (with general stationary arrival process) is just the arrival rate divided by the individual service rate [e.g., see (4.2.3) of Ref. 19], eq. (28) would be valid in the original closed model if (1) $\underline{\lambda}$ were the vector of arrival rates to the IS node and (2) $\sum_{i=1}^{q} EN_{ij}^o(\underline{\lambda})$ were the correct mean number of class j customers in the q-node subnetwork. However, $\sum_{i=1}^{q} EN_{ij}^o(\underline{\lambda})$ is in fact an approximation based on both $\underline{\lambda}$ and the Poisson assumption. Even if $\underline{\lambda}$ were correct, $EN_{ij}^o(\underline{\lambda})$ would be an approximation. In Section VIII we show that both conditions are satisfied asymptotically if we let $K_j \to \infty$, $\mu_j \to 0$, and $K_j\mu_j \to \lambda_j$ for each j. (This is completely established only for the case of one job class, but we conjecture that the convergence is valid for multiple job classes too.) Hence, there is reason to expect that the procedure will perform well as an approximation for the closed model under certain conditions. Interestingly, as indicated in Section I, these conditions are the same as those in Ref. 14.

Equation (28) also coincides with FPM method. With the FPM method we approximately solve the original closed model by finding the external arrival rate in the associated $(q + 1)$-node open network that makes the expected equilibrium population precisely K_j for class j. (We now regard the IS node as part of the open network.) However, the expected population of class j customers in the IS node is λ_j/μ_j, so that (28) is equivalent to

$$K_j = \sum_{i=1}^{q+1} EN_{ij}^{o}(\lambda) \tag{29}$$

for each j.

To complete the specification of the FPM procedure here, we must identify the vector λ satisfying (28) for all j. This can usually be done iteratively. The key is to recognize that the vector $\{EN_{ij}^{o}(\lambda), 1 \le i \le q, 1 \le j \le p\}$ is a strictly increasing continuous function of $(\lambda_1, \cdots, \lambda_p)$. We first bound λ_j above by

$$U_j^{(1)} = K_j \mu_j, \ 1 \le j \le p. \tag{30}$$

Then we bound λ_j below and above successively by

$$L_j^{(k)} = \left(K_j - \sum_{i=1}^{q} EN_{ij}^{o}(U_1^{(k)}, \cdots, U_p^{(k)})\right)\mu_j$$

$$U_j^{(k+1)} = \left(K_j - \sum_{i=1}^{q} EN_{ij}^{o}(L_1^{(k)}, \cdots, L_p^{(k)})\right)\mu_j \tag{31}$$

for $k \ge 1$, $1 \le j \le p$. It is easy to see that

$$L_j^{(k)} < L_j^{(k+1)} < \lambda_j < U_j^{(k+1)} < U_j^{(k)}, \tag{32}$$

$L_j^{(k)} \to \lambda_j$, and $U_j^{(k)} \to \lambda_j$ for λ_j satisfying (28). [Use the fact that $EN_{ij}^{o}(\lambda)$ is a continuous strictly increasing function of λ.]

To properly initialize the procedure, we must of course have (30) be feasible arrival rates; i.e., we need stability:

$$EN_{ij}^{o}(U_1^{(1)}, \cdots, U_p^{(1)}) < \infty \tag{33}$$

for all i and j. Moreover, we need

$$\sum_{i=1}^{q} EN_{ij}^{o}(U_1^{(1)}, \cdots, U_p^{(1)}) < K_j \tag{34}$$

for each j to have (31) be feasible rates. Conditions (33) and (34) should hold if indeed K_j is large, μ_j is small, and $K_j\mu_j$ is not too large. If conditions (33) and (34) were violated, we could search for initial conditions satisfying the appropriate monotonicity.

In fact, such elaborate analysis as we have just described is often unnecessary. If, indeed, K_j is large and μ_j is small for each j, with no node in the q-node network in heavy traffic, then it often suffices to use the simple formula (30) as the approximation for the external arrival rate λ_j. As indicated in (32), the approximation (30) yields upper bounds for λ_j and $EN_{ij}^{o}(\lambda)$ for all i and j. Alternatively, the first lower bound in (32) is often a good approximation, see Section VII.

It is intuitively obvious that the first upper bound (30) for λ in (28)

is also an upper bound for the vector of throughputs in the original closed model, and we prove this in Section VIII. It also seems plausible that the equilibrium queue length vector $\{N_{ij}^o(\lambda)\}$ in the open model with λ in (30) would be a stochastic upper bound for the corresponding random vector in the closed model, and we also prove this for the case of a single job class. (The more general case of multiple job classes remains a conjecture.) Theorem 1 in Section IV establishes for a single job class that the FPM throughput in (28) is a lower bound for the throughput in the original closed model. This extends to multiple job classes by the remark following Theorem 4. Hence, for each job class j, we have

$$L_j^{(1)} \le L_j^{(k)} \le \lambda_j \equiv \theta_j^o \le \theta_j^c \le U_j^{(1)}. \tag{35}$$

VII. EXAMPLES WITH A DECOUPLING INFINITE-SERVER NODE

In this section, we illustrate the FPM method in Section VI by returning to Example 3 in Section 1.4, which is the central processor model treated by McKenna, Mitra, and Ramakrishnan.[13] This is a closed cyclic product-form network with two nodes. The first node is the CPU, where service is provided according to the processor-sharing discipline. Equivalently for our purposes because of insensitivity properties,[5,6,19,20] the service discipline can be FCFS with an exponential service-time distribution. The second node is a think node, which is an IS node, representing independent delays at terminals before a job is next sent to the CPU. Again, because of insensitivity properties, only the mean of the service-time distribution at the IS node matters. We shall first consider the case of one job class, which is test problem 2 described in Table I of Ref. 13. Then we consider two job classes and finally we consider the special case of a population of size one to show that the FPM method can perform poorly when the approximating conditions do not nearly hold.

7.1 One job class

The specified model in Ref. 13 is closed with a fixed population size (also referred to as degree of multiprogramming). We shall consider the associated open model obtained by cutting the arrivals to node 1 and inserting an external Poisson arrival process with rate λ. This open model has a very simple solution. To express it, let μ_1^{-1} be the mean processing time for each job at the CPU and μ_2^{-1} the mean think time (individual service time at node 2). Let $\rho_1 = \lambda/\mu_1$ and $\alpha_2 = \lambda/\mu_2$ and assume that $\rho_1 < 1$ to guarantee stability. The equilibrium distribution in the open model has independent marginal distributions with the marginal being geometric at node 1 and Poisson at node 2:

$$P(N_1^o = k_1, N_2^o = k_2) = (1 - \rho_1)\rho_1^{k_1} e^{-\alpha_2}\alpha_2^{k_2}/k_2! \tag{36}$$

With the FPM method, we set the expected equilibrium total population equal to K; i.e.,

$$EN^o = \frac{\rho_1}{1-\rho_1} + \alpha_2 = \frac{\lambda}{\mu_1 - \lambda} + \frac{\lambda}{\mu_2} = K. \tag{37}$$

Hence, we obtain the following formula for the external arrival rate λ:

$$2\lambda\mu_1^{-1} = 1 + x(1+K) - \sqrt{[1+(1+K)x]^2 - 4Kx}, \tag{38}$$

where $x = \mu_2/\mu_1$. By using a Taylor series expansion in powers of x, we see that

$$\lambda/\mu_1 = Kx - Kx^2 - K(K-1)x^3 + 0(x^3), \tag{39}$$

so that $\lambda \approx K\mu_2$ when μ_2 and $\mu_2 K$ are sufficiently small compared to μ_1.

In contrast, for the closed model we combine (9) and (36). This is conceptually simple, but the calculation can be complicated for large population sizes. McKenna, Mitra, and Ramakrishnan[13] used this example to illustrate the advantage of PANACEA over previous convolution algorithms for the closed model, such as are contained in the software package CADS.[21] For large population sizes, CADS was unable to obtain a solution, while PANACEA obtained a solution easily. Moreover, for all population sizes, the throughputs calculated by the two methods agree closely.

It is significant that comparable results can be obtained for this example by the FPM method by hand. We do not even need to use (38); we can simply use (30) to obtain $\lambda = K\mu_2$. A comparison of the throughput calculations appears in Table I. In this example we have small μ_2 ($\mu_2 = 1/240$ and $\mu_1 = 1$) and large K ($K = 10$, 50, 100, and 200). Since the FPM method throughput provides a lower bound on the closed network throughput, the FPM answer is essentially exact for $K \leq 100$. As in Refs. 13 and 14, the FPM procedure works best here if μ_2 is small, K is large, and $K\mu_2$ is not too close to μ_1. Under heavy loads, the open-network M/M/1 formula at node 1 keeps the throughput down with the FPM method.

The last two columns of Table I contain the first two and first three terms of the Taylor series expansion in (39); the first term of course corresponds to the first upper bound in (30), which appears earlier in the table. Evidently the algorithm in Section VI converges faster than the Taylor series for larger values of K.

7.2 Two job classes

We now give additional details for test problem 2 in Ref. 13, which differs from Test Problem 1 only by having two job classes. Node 1 (the CPU) is again a processor-sharing node and node 2 is the IS node.

The mean service times for the two classes are 1 and 1.5 at the CPU and 450 and 150 at the IS node, respectively. Let μ_{ij} be the service rate of class j at node i and let K_j be the population of class j. The first upper bounds for the approximate arrival rate, obtained from (30) are

$$U_j^{(1)} = K_j\mu_{2j}. \tag{40}$$

The associated approximate CPU utilization of class j, say ρ_j, is thus $\rho_j \approx K_j\mu_{2j}/\mu_{1j}$ and the associated approximate total utilization of the CPU, say ρ, is $\rho \approx \rho_1 + \rho_2$. The first lower bounds on the approximate arrival rates, obtained from (31), are

$$L_j^{(1)} = (K_j - \rho_j/(1 - \rho))\mu_{2j}. \tag{41}$$

The associated approximation for the total CPU utilization is thus

$$\rho \approx L_1^{(1)}/\mu_{11} + L_2^{(1)}/\mu_{12}. \tag{42}$$

In this case we only compute the first upper and lower bounds. Table IV shows that the approximation procedure works very well. We are able to produce results very close to those given in Ref. 13 by hand in a few minutes. As suggested in Section VI, the first lower bound in (31) and (41) seems to provide a good approximation, even though it is a lower bound. From (35), the first upper and lower bounds from (40) and (41) are upper and lower bounds on the throughput in the closed model.

7.3 Where the FPM method performs poorly

When the populations are not large, the FPM method can perform poorly. This is easily and dramatically demonstrated with the same two-node closed network in which there is a single job. Let node 1 have one exponential server at rate 1 and let node 2 be the designated IS node with individual service rate x.

The exact equilibrium distribution has the job at node 1 with probability $x/(1 + x)$, which is also the associated long-run flow rate

Table IV—A comparison of the approximation method with exact results for the two-class example in Section 7.2

Number of Jobs (Degree of Multiprogramming) Class 1/Class 2	Total Utilization of CPU				
	For the Closed Model			New Approximation	
	From Ref. 13		Version 2.1 of PANACEA	First Upper Bound (40)	First Lower Bound (41)
	CADS	PANACEA			
10/10	0.118	0.119	0.121	0.122	0.121
50/50	0.593	0.60	0.599	0.611	0.598
100/50	Breakdown	0.69	0.706	0.722	0.704
200/10	Breakdown	0.54	0.540	0.544	0.540

out of node 2. However, the actual arrival process to node 1 is a renewal process in which the renewal interval is the sum of two independent exponential variables, one with mean 1 and the other with mean $1/x$. Moreover, the arrival rate depends dramatically on the state.

When we carry out the approximation procedure, we treat node 1 as an M/M/1 queue, so that (28) becomes

$$\lambda = [1 - \lambda/(1 - \lambda)]x, \tag{43}$$

which requires $\lambda < 0.5$ to have a solution. As $x \to \infty$, $\lambda(x) \to 1/2$. Obviously, the approximation does not work well in this case. The approximate throughput approaches one-half, while the true value in the closed model approaches 1 as $x \to \infty$. The normalized difference $\Delta = (\theta^c - \theta^o)/\mu_1$ approaches one-half, which in Section III we conjectured was the lower bound.

VIII. SUPPORTING THEORY WITH AN INFINITE-SERVER NODE

In this section we establish some theoretical results that help explain why and when the FPM algorithm in Section VI approximates the closed models well. As in Section VI, we assume that there is an IS node visited by all classes. We show that the subnetwork of the closed Markov network without the IS node approaches an open Markov network as the populations increase and the service rates at the IS node decrease appropriately (see Theorems 8 and 9). As a consequence, we show that the FPM method is asymptotically correct for the closed model under these conditions (see Theorem 12).

8.1 A sequence of closed models

As in Section VI, there are p job classes and $q + 1$ nodes with node $q + 1$ being the IS node that is visited by every class. We consider a sequence of systems indexed by the superscript n. Let μ_j^n be the individual service rate of class j at node $q + 1$ in the nth system. Let K_j^n be the fixed customer population of class j in system n. As with Poisson approximations for the binomial distribution[41] and as in Ref. 18, the idea is to let $K_j^n \to \infty$ and $\mu_j^n \to 0$ in such a way that $K_j^n \mu_j^n \to \lambda_j$ for each j as $n \to \infty$.

We let the remaining network structure and parameters be fixed, independent of n; neither the total numbers of nodes $q + 1$ and classes p, nor the parameters of the q-node subnetwork change with n. We still assume the basic Markov Jackson network structure specified in Section I, modified to allow multiple classes, but many of the results extend to more general models (see subsequent remarks).

Let p_{ji} be the probability that a departure of class j from the IS node goes next to node i in the q-node subnetwork. (There could be imme-

diate feedback to the IS node, which occurs for class j with probability $1 - \sum_{i=1}^{q} p_{ji}$.)

Let $A_{ji}^{cn}(t)$ be the counting process in the nth closed system representing the number of departures of class j from the node $q + 1$ in the interval $[0, t]$ that go next to node i. Let $N_{ij}^{cn}(t)$ represent the number of class j customers at node i at time t in the nth closed system. Let $\underset{\sim}{A}_{ji}^{cn}$, $\underset{\sim}{A}^{cn}$, $\underset{\sim}{N}_{ij}^{cn}$, and $\underset{\sim}{N}^{cn}$ represent the associated stochastic processes, i.e.,

$$\underset{\sim}{A}_{ji}^{cn} \equiv \{A_{ji}^{cn}(t), t \geq 0\}$$

$$\underset{\sim}{A}^{cn} \equiv \{\underset{\sim}{A}_{ji}^{cn}; 1 \leq j \leq p, 1 \leq i \leq q\}$$

$$\underset{\sim}{N}_{ij}^{cn} \equiv \{N_{ij}^{cn}(t), t \geq 0\}$$

$$\underset{\sim}{N}^{cn} \equiv \{\underset{\sim}{N}_{ij}^{cn}; 1 \leq i \leq q, 1 \leq j \leq p\}. \tag{44}$$

We can initialize the closed networks at time 0 in various ways. For example, we could assume that all $K_1^n + \cdots + K_p^n$ customers initially are at node $q + 1$. We will later simply assume that the initial distributions converge to a proper limit, which includes this situation as a special case.

Let $\Pi_{ji}(\lambda_{ji}) \equiv \Pi_{ji}(\lambda_{ji}, t) \equiv \{\Pi_{ji}(\lambda_{ji}, t), t \geq 0\}$ be a Poisson counting process with intensity λ_{ji}, and let $\underset{\sim}{\Pi} \equiv \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ be a pq-dimensional vector of independent Poisson processes with intensities $\underset{\sim}{\lambda} \equiv (\lambda_{ji}; 1 \leq j \leq p, 1 \leq i \leq q)$, i.e.,

$$\underset{\sim}{\Pi}(\underset{\sim}{\lambda}) \equiv \{\Pi_{ji}(\lambda_{ji}), 1 \leq j \leq p, 1 \leq i \leq q\}. \tag{45}$$

Let $N_{ij}^{o}(t)$ represent the number of class j customers at node i at time t in the q-node open network obtained by deleting the IS node and replacing its departure process with the external Poisson arrival process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$. By "external" we mean that we have the standard open model in which future arrivals are independent of the network state and history; i.e., $\{\Pi_{ji}(\lambda_{ji}, t + u) - \Pi_{ji}(\lambda_{ji}, t), u \geq 0, 1 \leq i \leq q, 1 \leq j \leq p\}$ is independent of $\{N_{ij}^{o}(s), s \leq t, 1 \leq i \leq q, 1 \leq j \leq p\}$ for each t. In both the open and closed models, successive service times and routings are mutually independent and independent of the history of the network prior to their generation. (At this point we are not using the FPM method in the open model; the arrival rates are simply specified as $\underset{\sim}{\lambda}$.) Let $\underset{\sim}{N}^{o}$ be the associated vector-valued stochastic process.

The following theorem expresses how the q-node subnetwork of the $(q + 1)$-node closed network without the IS node approaches a q-node open network as $n \to \infty$. The convergence of stochastic processes described below is convergence in distribution (weak convergence),

which we denote by $\Rightarrow$ (see Refs. 28 and 29 and references there). The stochastic processes are random elements of the function space $D[0, \infty) \equiv D([0, \infty), R^{pq})$.

Theorem 8: Let $\lambda_{ji} = \lambda_j p_{ji}$ *for each* j *and* i. *If* $K_j^n \to \infty$, $\mu_j^n \to 0$, $K_j^n \mu_j^n \to \lambda_j$ *for each* j, *and* $\underset{\sim}{N}^{cn}(0) \Rightarrow \underset{\sim}{N}^{o}(0)$ *in* R^{pq} *as* $n \to \infty$, *where* $\underset{\sim}{N}^{o}(0)$ *is a proper random vector* $[P(N_{ij}^{o}(0) < \infty) = 1$ *for all* i *and* $j]$, *then as* $n \to \infty$

$$(a)\ \underset{\sim}{A}^{cn} \Rightarrow \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$$

and

$$(b)\ \underset{\sim}{N}^{cn} \Rightarrow \underset{\sim}{N}^{o}.$$

(c) *If, in addition,* $\{[N_{ij}^{cn}(0)]^k\}$ *is uniformly integrable, then*

$$E[N_{ij}^{cn}(t)]^k \to E[N_{ij}^{o}(t)]^k$$

for each i, j, *and* t.

In the proof, as in Section IV, we use the notion of stochastic order. One random element (random element of R, R^k, $D[0, \infty)$, etc.) X_1 is stochastically less than or equal to another X_2, denoted by $X_1 \leq_{st} X_2$, if $Eh(X_1) \leq Eh(X_2)$ for all nondecreasing real-valued functions h for which the expectations are well defined.[33] For this a partial ordering must be defined on the sample space, which we take to be the usual one; e.g., $(x_1, \cdots, x_k) \leq (y_1, \cdots, y_k)$ in R^k if $x_i \leq y_i$ for each i and $\{x(t), t \geq 0\} \leq \{y(t), t \geq 0\}$ in $D[0, \infty)$ if $x(t) \leq y(t)$ for each t.

Proof: (a) The proof follows Ref. 18, which establishes convergence to a Poisson process for the departure process of certain G/GI/∞ queues under similar conditions. The result is not already contained in Ref. 18 because the arrival process to the IS node here is changing with n. However, by Corollary 1 to Theorem 1 in Ref. 18, it suffices to show that $\underset{\sim}{\Lambda}^{cn} \Rightarrow \underset{\sim}{\lambda}\,\omega$, where $\omega(t) = 1$, $t \geq 0$, as in (3.1) of Ref. 18, and $\underset{\sim}{\Lambda}^{cn}$ is the stochastic intensity of the counting process $\underset{\sim}{A}^{cn}$, defined by

$$\Lambda_{ji}^{cn}(t) = [K_j^n - N_{ij}^{cn}(t)]\mu_j^n p_{ji}, \qquad t \geq 0. \tag{46}$$

(For related theory, see Ref. 30 and references there.) The desired weak convergence of $\underset{\sim}{\Lambda}^{cn}$ follows easily from (46) because, for any T,

$$\sup_{0 \leq t \leq T} N_{ij}^{cn}(t) \leq_{st} N_{ij}^{cn}(0) + \Pi_{ji}(K_j^n \mu_j^n p_{ji}, T), \tag{47}$$

where $\leq_{st}$ denotes stochastic order defined above and the two quantities on the right are independent. Since we have assumed that $\underset{\sim}{N}^{cn}(0)$ converges and that $N_{ij}^{o}(0)$ is proper and since $K_j^n \mu_j^n p_{ji} \to \lambda_{ji}$ as $n \to \infty$, (47) implies that the sequence $\{\underset{\sim}{N}^{cn}\}$ is uniformly tight (see p. 37 of Ref. 28). This implies that $\underset{\sim}{N}_{ij}^{cn} \mu_j^n \Rightarrow 0\omega$ as $n \to \infty$ and the desired conclusion.

(b) Convergence in distribution of $\underset{\sim}{N}^{cn}$ follows by model continuity

as in Refs. 31 and 32. In particular, given part (a), we can construct versions of $\underline{A}^{cn}$ and $\underline{\Pi}(\underline{\lambda})$ on the same sample space so that there is convergence of the sample paths, using the Skorohod embedding theorem.[29] Using the same service times and routing in all systems, we obtain convergence of the sample paths of $\underline{N}^{cn}$ to $\underline{N}^{o}$ with probability one on the specially constructed space. (Since $\underline{\Pi}(\underline{\lambda})$ has no fixed jump points, simultaneous transitions need not be considered.) This implies convergence in distribution of the separate stochastic processes.

(c) The stochastic dominance used in part (a) and the new condition imply that the random variables $\{[N_{ij}^{cn}(t)]^k, n \geq 1\}$ are uniformly integrable (see p. 32 of Ref. 28). Part (b) implies that $N_{ij}^{cn}(t) \Rightarrow N_{ij}^{o}(t)$ as $n \to \infty$ for each i, j, and t. Theorem 5.4 of Ref. 28 thus implies convergence of the moments. □

Remarks: 1. All the conditions on $N_{ij}^{cn}(0)$ hold trivially if all $K_1^n + \cdots + K_p^n$ jobs are initially at the IS node for each n.

2. The conditions in Theorem 8 can be relaxed. The q-node subnetwork of the closed model can be quite general; e.g., the service-time distributions can be nonexponential with FCFS nodes. Our proof only exploits the fact that the service-time distribution at the IS node is exponential. The service-time distribution at the IS node could be made general too, as in Ref. 18, but then we would have to be careful with the initial conditions. If the initial residual service-time distributions at the servers are independent stationary-excess distributions of a service-time distribution that has no mass at zero, then part (a) holds by virtue of the limit theorem for the superposition of independent and identically distributed (i.i.d.) stationary renewal processes.[51] Of course, if the service-time distribution has positive mass at zero, then the limit process is, instead, batch Poisson with geometric batches. □

Theorem 8 implies that the random vectors $\underline{A}^{cn}(t)$ and $\underline{N}^{cn}(t)$ in R^{pq} converge in distribution as $n \to \infty$ for each t, but Theorem 8 says nothing about the equilibrium distributions. In fact, we have not yet ruled out the possibility that the open network is unstable; i.e., we could have $N_{ij}^{o}(t) \Rightarrow \infty$ as $t \to \infty$. Indeed, Theorem 8 is still valid in this case, but now we consider the equilibrium distributions. Let $\underline{N}^{cn}(\infty)$ and $\underline{N}^{o}(\infty)$ be random vectors with the equilibrium or limiting distributions as $t \to \infty$. (For the continuous-time Markov chains, they are necessarily unique.) We assume that the limiting Poisson intensities λ_{ij} are small enough so that the equilibrium or limiting distribution for $\underline{N}^{o}(t)$ exists.

Theorem 9: Assume that a proper equilibrium distribution exists for $\underline{N}^{o}$. Also, assume that either (1) there is a single job class or (2) the

sequence $\{\underset{\sim}{N}^{cn}(\infty), n \geq 1\}$ *is uniformly tight. Then, under the conditions of Theorem 8,* $\underset{\sim}{N}^{cn}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ *in* R^{pq} *as* $n \to \infty$.

We defer the proof of Theorem 9 until we develop some stochastic comparison tools, which are interesting in their own right. We are able to establish the desired stochastic comparison result (see Theorem 11) only when there is a single job class, which explains the second assumption in Theorem 9. We conjecture that the required tightness in Theorem 9 can be proved from the other assumptions for multiple classes.

8.2 Stochastic comparisons

For the counting processes $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$, we use the notion of stochastic order based on conditional failure rates or stochastic intensities, introduced in Ref. 34. The stochastic intensity of the vector-valued stochastic process $\underset{\sim}{A}^{cn}(t) \equiv \{A_{ji}^{cn}(t)\}$ is defined in (46). Of course, the stochastic intensity of the Poisson process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ is the deterministic function $\underset{\sim}{\lambda}\omega$. Following Ref. 34, the counting process $\underset{\sim}{A}^{cn}$ is said to be stochastically less than or equal to the Poisson process $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ in the sense of conditional failure rates, here denoted by $\underset{\sim}{A}^{cn} \leq_f \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$, if

$$\Lambda_{ji}^{cn}(t) \leq \lambda_{ji} \tag{48}$$

with probability 1 for all j, i, and t ($\leq_1$ is used in Ref. 34). From (46), it is easy to see that indeed (48) is satisfied. Hence, trivially we have Theorem 10.

Theorem 10: If $K_j^n \mu_j^n \leq \lambda_j$ *for each* j, *then* $\underset{\sim}{A}^{cn} \leq_f \underset{\sim}{\Pi}(\underset{\sim}{\lambda})$.

Corollary to Theorem 9: In the setting of Section VI, $K_j\mu_j$ *is an upper bound for the expected average throughput for class* j *over any time interval. Hence, the first upper bound for the FPM method in (30) yields an upper bound for the long-run throughput of each class in the closed method.*

We now establish a general stochastic comparison between $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$. We exploit a coupling or special, almost surely ordered construction, as in Refs. 33 and 34. To establish a general comparison result for $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$, we assume that there is a single customer class. We thus drop the j subscript. We also exploit the fact that the processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ are continuous-time Markov processes, but now the service rate at node i when there are k customers present can be a general nondecreasing function, say $\mu_i(k)$, for $1 \leq i \leq q$.

Theorem 11: Suppose that there is a single job class with $K^n\mu^n \leq \lambda$. *Let the processes* $\underset{\sim}{N}^{cn}$ *and* $\underset{\sim}{N}^{o}$ *be Markov with the service rate functions* $\mu_i(k)$ *nondecreasing in* k *for each* i, $1 \leq i \leq q$.

(a) *If* $\underset{\sim}{N}^{cn}(0) \leq_{st} \underset{\sim}{N}^{o}(0)$ *in* R^q, *then* $\underset{\sim}{N}^{cn} \leq_{st} \underset{\sim}{N}^{o}$ *in* $D[0, \infty)$.

(b) *If, in addition, the equilibrium distribution for the open network exists, then also* $\underset{\sim}{N}^{cn}(\infty) \leq_{st} \underset{\sim}{N}^{o}(\infty)$ *in* R^q.

Proof: (a) The argument parallels that of Theorems 6, 7, and 10 in Ref. 34. For more details on the method, see Sonderman.[53] First, Theorem 10 implies that versions of the arrival processors $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{\Pi}(\underset{\sim}{\lambda})$ can be constructed on the same probability space so that the points of $A_{ji}^{cn}(t)$ form a subsequence of the points in $\Pi_{ji}(\lambda_{ji}, t)$ for each j ($j = 1$ here) and i (the ordering $\leq_2$ in Ref. 34). Next we can construct the service completions for $\underset{\sim}{N}^{cn}$ using the service completions of $\underset{\sim}{N}^{o}$. If there is a service completion at node i in process N^o at time t, then we let there be a corresponding service completion at node i in $\underset{\sim}{N}^{cn}$ with probability $\mu_i(N_i^{cn}(t))/\mu_i(N_i^o(t))$. When there are service completions in both processes, we let the routing be identical. By using induction on the transition epochs, we see that this special construction keeps the sample paths ordered and the distributions of the individual stochastic processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ unchanged.

(b) The stochastic order for each t as a consequence of part (a) is preserved in the limit as $t \to \infty$ (see Proposition 3 of Ref. 33). □

Remarks: 1. It is not difficult to see that Theorem 11(a) is not true for multiple job classes. For example, consider a network with two nodes plus the IS node and three job classes. Let class j jobs go from the IS node to node j and then back to the IS node for $j = 1, 2$. Let class 3 jobs go from the IS node to node 2, then node 1 and then back to the IS node. Let all service rates be identical at nodes 1 and 2. Let K_1^n and K_3^n be large and μ_1^n and μ_3^n be small so that the arrival processes of classes 1 and 3 are both nearly Poisson in the closed model. On the other hand, let $K_2^n = 1$, so that $\underset{\sim}{A}_{22}^{cn}$ is considerably smaller (stochastically) than the Poisson process associated with the open model. Let nodes 1 and 2 be initially empty. For some relatively short initial time interval, say $[0, t]$, in the open model there are more arrivals of class 2 to node 2, with negligible change for classes 1 and 3. These class 2 jobs at node 2 tend to impede the class 3 jobs at node 2, so that the class 3 jobs come to node 1 more slowly in the open model. Hence, the class 1 jobs can get through node 1 more easily; thus, we can have $EN_{11}^o(t) \leq EN_{11}^{cn}(t)$ even though $K_1\mu_1 \leq \lambda_1$.

2. Even though Theorem 11(a) does not extend to multiple job classes, we conjecture that Theorem 11(b) does. That would be sufficient to eliminate conditions (1) and (2) in Theorem 9.

Proof of Theorem 9: Since $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ are continuous-time Markov processes with the given equilibrium distributions, we can apply Theorem 8(b) here and Lemma 1 of Ref. 31. This implies that the desired convergence $\underset{\sim}{N}^{cn}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ holds provided that $\{\underset{\sim}{N}^{cn}(\infty), n \geq 1\}$ is uniformly tight. We use the fact that $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ have unique equilibrium distributions. For the case of a single job class, the sequence

$\{\underset{\sim}{N}^{cn}(\infty), n \geq 1\}$ is uniformly tight by Theorem 11(b). The stochastic dominance implies the desired uniform tightness because each individual probability measure is tight (see Theorem 1.4 of Ref. 28). □.

Remarks: 1. Of course, Theorem 9 applies to other non-Markov product-form models that have the same equilibrium distributions by virtue of insensitivity properties.[3-6,19,20]

2. Theorem 9 also holds for more general service-time distributions in the q-node subnetwork provided that we can establish the uniform tightness. The original processes $\underset{\sim}{N}^{cn}$ and $\underset{\sim}{N}^{o}$ can be made Markov by appending supplementary variables.

3. If the service-time distribution for class j at the IS node is phase type instead of exponential, then Theorem 10 remains valid with $\lambda_{ji}^{*} \geq K_j^n \mu_j^{*n}$, where μ_j^{*n} is the maximum phase service rate for class j. If the open network process $\underset{\sim}{N}^{o}$ is stable with the high intensities $\underset{\sim}{\lambda}^{*}$, then we can apply the analog of Theorem 11 to obtain the tightness needed in Theorem 9 (again for a single job class).

We now show that the first lower bound for the FPM method in (31) is a lower bound for the throughput in the closed network. As stated, this follows from (32) and Theorem 1, but we make stronger comparisons using the stochastic intensity $\underset{\sim}{\Lambda}^{cn}$ of the arrival process $\underset{\sim}{A}^{cn}$, defined in (46).

Corollary to Theorem 11: Suppose that there is a single job class with $K^n \mu^n \leq \lambda$. Then, for each i and t,

$$(a)\ \Lambda_{1i}^{cn}(t) \geq_{st} [K^n - N^o(t)]\mu^n p_{1i}$$

and

$$(b)\ E\Lambda_{1i}^{cn}(t) \geq E[K^n - N^o(t)]\mu^n p_{1i}.$$

If, in addition, both systems are in equilibrium, then

$$(c)\ E\left\{t^{-1}\int_s^{s+t} \Lambda_{1i}^{cn}(u)du\right\} \geq L_1^{(1)} \quad \text{for all } s \text{ and } t$$

and

$$(d)\ \theta^c \geq L_1^{(1)}.$$

8.3 The FPM method is asymptotically correct

We now apply Theorems 8 and 9 to deduce that the FPM method in Section VI is asymptotically correct. Due to the second assumption in Theorem 9, we only completely treat the case of one job class. Let $\underset{\sim}{A}^{on}$ and $\underset{\sim}{N}^{on}$ be the vector-valued arrival process and queue length process obtained by using the FPM method with the nth closed model.

Theorem 12: Under the conditions of Theorem 8, as $n \to \infty$,

$$(a)\ \underset{\sim}{A}^{on} \Rightarrow \underset{\sim}{\Pi}(\underset{\sim}{\lambda}),$$

$$(b)\ \underset{\sim}{N}^{on} \Rightarrow \underset{\sim}{N}^{o},$$

and

$$(c)\ E(N_{ij}^{on}(t))^k \to E(N_{ij}^{o}(t))^k$$

for each i, j, k, and t.

(d) Under the conditions of Theorem 9, for sufficiently large n, $\underset{\sim}{N}^{on}(\infty)$ *exists as a proper random vector and* $\underset{\sim}{N}^{on}(\infty) \Rightarrow \underset{\sim}{N}^{o}(\infty)$ *in* R^{pq}.

Proof: (a) Since $\underset{\sim}{A}^{on}$ is a Poisson process for each n, it suffices to show that the associated arrival rates converge. For this, it suffices to show that the difference between the first lower bound in (31) and the upper bound in (30) is asymptotically negligible, which is immediate under the conditions of Theorem 8. Parts (b) and (c) follow exactly as in Theorem 8. Theorems 10 and 11 extend easily when $\underset{\sim}{A}^{on}$ and $\underset{\sim}{N}^{on}$ replace $\underset{\sim}{A}^{cn}$ and $\underset{\sim}{N}^{cn}$ since the limiting system is the first upper bound for the FPM method. Finally, part (d) follows exactly as in Theorem 9. □

IX. A BOTTLENECK NODE WITH A LARGE POPULATION

9.1 A different approximation procedure

In this section we observe that the methods and results of Sections VI through VIII also apply, after appropriate modification, to closed networks with a bottleneck non-IS node. We first consider the case of one job class. For large populations, all servers at the bottleneck node will usually be busy, so that we can approximately analyze the original closed model by using the bottleneck node to decouple the network just as we used the IS node in Section VI. We remove the bottleneck node and replace its departure process by an external arrival process. We then solve, exactly or approximately, the resulting open network. If there are s servers at the bottleneck node, then the external arrival process would be the superposition of s i.i.d. renewal processes each having the bottleneck service-time distribution as the renewal-interval distribution. The routing of the external arrivals is just the original routing from the bottleneck node. When the service-time distribution at the bottleneck node is exponential, the approximating external arrival process is thus Poisson.[51] Otherwise, we would approximately characterize the external superposition arrival process as in Ref. 7 and apply the algorithm there to approximately analyze the resulting non-Markov open network.

In this setting the bottleneck node is easy to identify. As in Section III, we begin by replacing one internal arrival process by the external arrival process with a rate sufficiently small to ensure stability. Let λ_i be the net arrival rate to node i obtained from solving the traffic rate

equations with the given external arrival rate, say λ_o. Then calculate the traffic intensity at node i as $\rho_i = \lambda_i/s_i\mu_i$ given that node i is a FCFS node with s_i servers, each working at rate μ_i. The node with the highest traffic intensity is the bottleneck node; call it node $q + 1$. We assume that there are no ties. The capacity of the network is thus $s_{q+1}\mu_{q+1}$. We can achieve any throughput less than $s_{q+1}\mu_{q+1}$ in the open model. The traffic intensity becomes 1 at node $q + 1$ at the capacity, which makes the system unstable. It of course is well known that $s_{q+1}\mu_{q+1}$ is an upper bound on the throughput even in non-Markov networks (see Ref. 11 and references there).

The proposed approximation procedure for the closed model with a large population is to solve the traffic rate equations for the associated open network and find the bottleneck node, which we denote by node $q + 1$. Then solve the open model obtained by deleting node $q + 1$ from the closed network and inserting an external arrival process with rate $s_{q+1}\mu_{q+1}$. However, unlike Sections III and VI, we do not use the FPM method for the full $(q + 1)$-node network; we do not require a consistency condition such as (28). We simply let the approximate number of jobs at node $q + 1$ in the original closed network be

$$EN^c_{q+1}(\infty) \approx K - \sum_{i=1}^{q} EN^o_i(s_{q+1}\mu_{q+1}). \tag{49}$$

If there are quite a few nodes but not a large population, we will use the original FPM method, but as the population grows with the number of nodes fixed, the effect of the bottleneck node becomes more pronounced.

9.2 *Limit theorems*

The approximation procedure just described is evidently quite well known. Supporting limit theorems are discussed by Whittle[35] and Brown and Pollett.[36] The methods and results of Section VIII provide a convenient way to prove that the approximation procedure is asymptotically correct for the q-node subnetwork excluding the bottleneck node as the population grows. Since the results and methods are similar to those in Section VIII, we only give a brief account. The analog of Theorem 8 is for one customer class. We let $K^n_1 \to \infty$ as before, but now we fix μ^n_1, the individual service rate at node $q + 1$. When the service-time distribution at the bottleneck node is exponential, we can use the obvious modification of the proof of Theorem 8(a). With general service-time distributions, it is easy to show that the probability that all servers are busy at the bottleneck node throughout any interval $[0, t]$ converges to 1 as $n \to \infty$. The rest of Section VIII applies in a straightforward manner, with essentially the same remarks about generalizations.

9.3 Multiple job classes

We now consider multiple job classes with a special bottleneck node. We assume that there is a single-server processor-sharing bottleneck node with fixed total service rate μ whenever any customers are present. Let the service requirements of class j at the bottleneck node be exponentially distributed with mean μ_j^{-1}. Let the population of class j in the network be K_j.

Again the approximation is obtained by replacing the bottleneck node by an external Poisson process with rate μ. Each of these external arrivals is from class j with probability

$$\gamma_j = K_j\mu_j/(K_1\mu_1 + \cdots + K_p\mu_p). \tag{50}$$

Consequently, as in Section VI, there is a pq-dimensional vector of independent Poisson processes with the intensity class j going to node i being $\lambda_{ji} = \mu\gamma_j p_{ji}$. The limit theorems in Section VIII also apply here. As before, Theorem 11(a) does not hold for multiple job classes.

9.4 Another stochastic comparison

We now make a stochastic comparison between the closed model and the open model resulting from the bottleneck approximation. We consider the case of one job class. We compare the q-dimensional equilibrium distribution of the subnetwork of the closed model without the bottleneck node to the q-dimensional equilibrium distribution in the q-node open model with external arrival rate $\mu_{q+1}s_{q+1}$. We show that the equilibrium distribution based on the bottleneck approximation is larger in a very strong sense, namely, in the MLR ordering used in Section 4.3.

Let $(N_1^c, \cdots, N_q^c)$ be the equilibrium random vector in the closed model with population K without the bottleneck node, defined in terms of an associated $(q + 1)$-dimensional open-model equilibrium random vector $(N_1^o, \cdots, N_{q+1}^o)$ by

$$P(N_1^c = k_1, \cdots, N_q^c = k_q) = \frac{P(N_1^o = k_1) \cdots P(N_q^o = k_q)P(N_{q+1}^o = K - \sum_{j=0}^{q} k_j)}{P(N^o = K)} \tag{51}$$

for $(k_1, \cdots, k_q)$ such that $k_1 + \cdots + k_q \leq K$.

Let $(N_1^b, \cdots, N_q^b)$ be the open-model equilibrium random vector with utilization at node i of u_i^o/u_{q+1}^o, where u_i^o is the utilization of node i in $(N_1^o, \cdots, N_{q+1}^o)$ in (51). We assume that $u_i^o < u_{q+1}^o$ for all i. This is tantamount to having an external Poisson arrival process with rate $u_{q+1}s_{q+1}$ in the q-node open network.

Theorem 13: $(N_1^c, \cdots, N_q^c) \leq_{lr} (N_1^b, \cdots, N_q^b)$.

Proof: It is immediate that the distribution of $(N_1^b, \cdots, N_q^b)$ is MTP_2 because the marginals are independent (see Proposition 3.5 of Ref. 50). Consequently, by Theorem 3 of Ref. 48, it suffices to show that $p_1(y)p_2(x) \leq p_1(x)p_2(y)$ for all $x \leq y$, where p_1 and p_2 are the associated probability mass functions. Moreover, it suffices to consider y differing from x by 1 in only one place, e.g., $x = (k_1, \cdots, k_q)$ and $y = (k_1 + 1, k_2, \cdots, k_q)$. We verify this as follows:

$$\frac{P(N_1^c = k_1 + 1, \cdots, N_2^c = k_2)\ P(N_1^b = k_1, \cdots, N_2^b = k_2)}{P(N_1^c = k_1, \cdots, N_2^c = k_2)\ P(N_1^b = k_1 + 1, \cdots, N_2^b = k_2)}$$

$$= \frac{P(N_1^o = k_1 + 1)P\left(N_{q+1}^o = K - \sum_{j=1}^{q} k_j - 1\right) P(N_1^b = k_1)}{P(N_1^o = k_1)P\left(N_{q+1}^o = K - \sum_{j=1}^{q} k_j\right) P(N_1^b = k_1 + 1)} \leq 1$$

because, for all j,

$$P(N_1^b = j + 1)/P(N_1^b = j) = u_{q+1}^o P(N_1^o = j + 1)/P(N_1^o = j)$$

and

$$P(N_{q+1}^o = j + 1)/P(N_{q+1}^o = j) \geq u_{q+1}^o. \tag{52}$$

To verify (52), recall that N_{q+1}^o has the equilibrium distribution of a birth-and-death process, so that

$$\hat{\lambda}_j P(N_{q+1}^o = j) = \hat{\mu}_{j+1} P(N_{q+1}^o = j + 1),$$

where $\hat{\lambda}_j$ is the arrival rate when $N_{q+1}^o = j$, which is independent of j, and $\hat{\mu}_{j+1}$ is the service rate when $N_{q+1}^o = j + 1$. When there is one server, $u_{q+1}^o = \hat{\lambda}_j/\hat{\mu}_{j+1}$ for all j, but in general, $\hat{\mu}_{j+1} = \mu_{q+1} \min\{j + 1, s\}$, so that $u_{q+1}^o \leq \hat{\lambda}_j/\hat{\mu}_{j+1}$. □

Remarks: 1. Theorem 13 has corollaries like those for Theorem 6. For example,

$$P(N_i^c \geq k_i \mid a_j \leq N_j^c \leq b_j) \leq P(N_i^b \geq k \mid a_j \leq N_j^b \leq b_j) \tag{53}$$

for all i, j, k_i, a_j, and b_j. Inequality (53) is interesting both when $i = j$ and $i \neq j$. Of course, when $i \neq j$, the right-hand side of (53) reduces to $P(N_i^b \geq k_i)$.

2. As in Section VIII, we can obtain results for the equilibrium distribution associated with other queue disciplines by invoking insensitivity properties.[3-6,19,20]

3. Algorithms for identifying bottleneck nodes and treating them are described by Schweitzer[54] and Goodman and Massey.[39] Stochastic bounds for open networks of single-server nodes are contained in

Massey.[38] These bounds apply to closed networks too by combining them with the comparison results in this paper.

4. As the population increases, the closed network can be said to be in heavy traffic. However, only the bottleneck node accumulates jobs in the limit. The number of jobs at the nonbottleneck nodes is asymptotically negligible compared to the number at the bottleneck node. In fact, by the analog of Theorem 9, the number of jobs at all nonbottleneck nodes, unnormalized, converges to a proper limit, as the population grows. Instead of the complicated multidimensional diffusion process approximations for networks of queues described in Reiman,[55] we have significant accumulation of customers only at the bottleneck node alone. The situation here is an example of the diffusion approximations with state space collapse discussed by Reiman.[56] However, because we are considering a closed network, the number of customers at the bottleneck node is best described by $K - \sum_{j=1}^{q} N_j^b$. Indeed, as a trivial corollary to the analog of Theorem 9, we have

$$(N_{q+1}^{cn} - K^n) \Rightarrow \sum_{j=1}^{q} N_j^b \tag{54}$$

as $n \to \infty$. Unless there are ties for the maximum traffic intensity, only one node will be a bottleneck node for both closed and open networks. Moreover, because of the geometric tails of the queue length equilibrium distributions in Markov networks, slight differences in traffic intensities will rapidly lead to large differences in the queue lengths as the population grows. Consequently, the case of a single bottleneck node treated here seems most relevant for applications.

X. APPROXIMATIONS FOR NON-MARKOV CLOSED NETWORKS

10.1 Several possible approximation procedures

Suppose, as in Ref. 7, that the Markov property is lost because we are considering FCFS nodes with nonexponential service-time distributions. There are two natural procedures for calculating approximate congestion measures for such non-Markov closed networks based on previously developed approximations for non-Markov open networks. Just as we can use Markov open models to analyze Markov closed models, we can use the approximate solution for an associated non-Markov open model to generate an approximate solution for the given non-Markov closed model.

The first procedure for non-Markov closed models starts with the approximate equilibrium distribution of the number of customers at each node in the associated open model, as described in Section III. Then the corresponding equilibrium distribution for the closed model can be obtained by conditioning as in (9). For the open model, the

standard approximation procedure is to use a product-form solution (an equilibrium distribution with independent marginal distributions). This is the procedure first suggested by Reiser and Kobayashi.[57] The Extended-Product-Form (EPF) method of Shum and Buzen[58,59] and the Generalized-Product-Form (GPF) method of Tripathi[60] are also variants of this approach. A complete approximation thus is determined by specifying the equilibrium distribution of the number of customers at each node in the open model. For example, with QNA[7] this can be done by fitting a discrete distribution to the quantities $P(N_i^? = 0)$, $E(N_i^?)$, and $\mathrm{Var}(N_i^?)$, which are currently provided in the model solution. In fact, in Ref. 7 an approximation for the waiting-time distribution at each node is obtained in this way. For single-server nodes, it is natural to use mixtures and convolutions of geometric distributions for the conditional distribution of the number of customers at each node, given that the server is busy. Such an approximation procedure based on QNA is currently being investigated.

There are some difficulties with this first procedure, however. We must do the same extensive calculation to find the normalization constant G as we do with the Markov closed model, so that we obtain no reduction in computation working with approximations. We can of course use many of the same algorithms now being used for Markov closed networks.[6]

The second procedure is to use the open model directly, as with the FPM method. We believe that this method can be expected to work about as well as it does for closed Markov models. Now, in the setting of Ref. 7 we also have variability parameters. In particular, we must specify a variability parameter as well as an arrival rate for the special new external arrival process.

There are three different situations. First, with a decoupling IS node containing most of the customers (under the conditions of Section VI), it is natural to use the FPM method and approximate the external arrival process by a Poisson process, so that there is no problem selecting the variability parameter; set it equal to 1. However, now it is important that the external Poisson arrival process replace the departure process from the special IS node. This process will be approximately Poisson, even with nonexponential service-time distributions.

Second, with a bottleneck node having s servers as in Section IX, it is natural to regard the arrival process as the superposition of s independent renewal processes each with the bottleneck service-time distribution as the renewal-interval distribution. When $s = 1$, the procedure is clear: use the squared coefficient of variation of the bottleneck service-time distribution. When $s > 1$, we can use approx-

imations for superposition processes as in Section 4.3 of Ref. 7. As described in Section IX, we would not use the FPM method, but instead the open model with the bottleneck node removed.

The third situation is where the FPM method is appropriate but the variability parameter needs to be determined. In Section 10.2 we discuss this case in detail.

There are of course many other procedures for approximately analyzing non-Markov closed networks with nonexponential FCFS nodes,[6,11,22,61-62] but we do not discuss them here.

10.2 The FPM method for non-Markov models

A simple procedure for the FPM method more generally, in the case of a single job class, is to first specify an external arrival rate λ_0 and then, for that specified arrival rate, solve a system of linear equations to obtain the variability parameter $c_0^2(\lambda_0)$ that makes the variability parameter of the departure process from the network equal to the variability parameter of the external arrival process. (The reason for doing this, of course, is that in the closed network these two processes are actually the same process.) We then solve the open model for a range of possible external arrival rates, associating $c_0^2(\lambda_0)$ with λ_0 each time. As before, the throughput when the population is K is the value of λ_0 such that $EN^o = K$.

We now describe in detail a modification of the QNA algorithm in Ref. 7 that has been developed to approximately analyze a closed non-Markov network of queues with one job class by the FPM method. The initial model is just as in Ref. 7 but without external arrival processes. In particular, the nodes have the FCFS discipline, several servers, and general service-time distributions. We assume that the reader is familiar with Ref. 7, and we use the same notation here.

The model input is a minor modification of the standard input in Section 2.1 of Ref. 7; we just omit the data for the external arrival processes. For each network we specify the following:

n = number of nodes in the network
m_j = number of servers at node j
τ_j = mean service time at node j
c_{sj}^2 = squared coefficient of variation of the service-time distribution at node j
q_{ij} = proportion of those customers completing service at node i that go next to node j.

To apply the FPM method, we introduce an external arrival process to one node, which we stipulate is node 1. To see how the expected network population depends on the external arrival rate, we specify a set of external arrival rates, which are understood to apply to node 1.

The set is specified by the following numbers:

L = lower bound for external arrival rate to node 1
U = upper bound for external arrival rate to node 1
C = number of different arrival rates.

Given the triple (L, U, C), the network will be analyzed C times with the following external arrival rates to node 1:

$$\lambda_{01} = L + k(U - L)/(C - 1) \tag{55}$$

for $k = 0, 1, \cdots, C - 1$. The external arrival rates to all other nodes are zero. To obtain the open model for the FPM method in each case, we insert an external arrival process to node 1 with one of the rates specified in (55) and we eliminate all internal arrivals to node 1. This is done with the algorithm by setting $q_{i1} = 0$ for all i.

We begin by solving the traffic-rate equations, given the external arrival rate λ_{01}, exactly as in Section 4.1 of Ref. 7. This provides the traffic intensities at the nodes, needed for the traffic variability equations.

Next we solve the traffic variability equations. The algorithm also determines the variability parameter c_{01}^2 for the external arrival process to node 1. As indicated above, the idea is to have the variability parameter of the external arrival process agree with the variability parameter of the total departure process from the network (which would have been the arrival process to node 1 in the closed model). The equations in (24) of Ref. 7 are valid for $j = 2, \cdots, n$; i.e., we have

$$c_{aj}^2 = a_j + \sum_{i=1}^{n} c_{ai}^2 b_{ij}, \qquad 2 \le j \le n, \tag{56}$$

with a_j and b_{ij} in (25) and (26) of Ref. 7. Since $q_{i1} = 0$ for all i, $c_{a1}^2 = c_{01}^2$. The variability parameters are solved by replacing the first equation in (24) of Ref. 7 with

$$c_{a1}^2 = \alpha_1^* + \sum_{i=1}^{n} c_{ai}^2 b_{i1}^*, \tag{57}$$

where

$$a_1^* = 1 + w_1^* \left\{ -1 + \sum_{i=1}^{n} (d_i/d) \left[\sum_{j=2}^{n} q_{ij} + \left(1 - \sum_{j=2}^{n} q_{ij}\right) \rho_i^2 \left(1 + \frac{c_{si}^2 - 1}{\quad}\right) \right] \right\}, \tag{58}$$

$$b_{i1}^* = w_1^* (d_i/d) \left(1 - \sum_{j=2}^{n} q_{ij}\right) (1 - \rho_i^2), \tag{59}$$

d_i is the departure rate from the network at node i, $d = d_1 + \cdots + d_n$ as in (23) of Ref. 7, and w_1^* is the superposition weighting function in (29) of Ref. 7 with p_{i1} in (30) there replaced by d_i/d.

We derive (57) through (59) as follows. First, the departure process from the whole network is the superposition of the departure processes (leaving the network) from the separate nodes. Hence, by Section 4.3 of Ref. 7,

$$c_{a1}^2 = w_1^* \left(\sum_{i=1}^{n} (d_i/d) c_{di}^{*2} \right) + 1 - w_1^*, \tag{60}$$

where c_{di}^{*2} is the variability parameter for the departure process from the network at node i; i.e.,

$$c_{di}^{*2} = \left(1 - \sum_{j=2}^{n} q_{ij} \right) c_{di}^2 + \sum_{j=2}^{n} q_{ij} \tag{61}$$

and

$$c_{di}^2 = 1 + (1 - \rho_i^2)(c_{ai}^2 - 1) + \frac{\rho_i^2}{\sqrt{m_i}} (c_{si}^2 - 1), \tag{62}$$

using first the splitting formula (36) and then the departure formula (39) from Ref. 7.

The rest of the modified QNA algorithm is just as in Ref. 7. We next calculate the congestion measures at the nodes using the traffic rate and variability parameters already determined. By running the algorithm a few times with various (L, U, C) triples, the user can easily select a set of external arrival rates to node 1 via (55) to yield a desired range of expected network populations in the open model. The algorithm also can automatically find the external arrival rate yielding a specified expected equilibrium network population.

We can also use the finite-waiting-room refinement introduced in Section 1.3 to calculate the congestion measures at the nodes in the open model. For single-server nodes, we use the modifications in (2) and (3), even if the nodes do not correspond to M/M/1 models.

Remarks. 1. Our procedure above replaces an internal arrival process to node 1 by an external arrival process. Instead, we could have replaced the internal departure process from node 1 by an external arrival process. It would then have been immediately split according to the routing probabilities q_{1i}. The original departures from node 1 would then be removed.

2. As mentioned in Section I, it may be desirable to artificially deflate the variability parameters of the arrival processes. It is natural to do this after the traffic variability equations have been solved as described above.

3. The procedure can easily be extended to multiple job classes in various ways. For example, we can let the variability parameters of the external arrival processes for each individual class be unspecified. We then can apply the procedure in (56) through (59) in this section to specify the variability parameter for the overall external arrival process in the aggregated single-class network obtained from Section 2.3 of Ref. 7. The only remaining complication is that instead of the external arrival rates λ_{0i} determined by the single triple (L, U, C), we now have a vector of external arrival rates determined by such a triple for each job class. Automatic search obviously becomes desirable in this setting.

4. The approximate solution using the FPM method can be fruitfully combined with the exact solution of the corresponding Markov model to obtain improved approximations for the closed non-Markov model. For example, we can solve the closed Markov model to obtain u_i^{cM} as the utilization of node i when the service-time distributions are all exponential. We can also apply the FPM method twice, once with general service-time distributions and once with exponential service-time distributions, to obtain corresponding utilizations u_i^{oG} and u_i^{oM}. We can then approximate u_i^{cG}, the utilization at node i in the closed network with nonexponential service-time distributions, by

$$u_i^{cG} = u_i^{cM} u_i^{oG} / u_i^{oM}. \tag{63}$$

Since (65) can lead to inconsistencies such as $u_i^{cG} > 1$, it is natural to use

$$(1 - u_i^{cG}) = (1 - u_i^{cM})(1 - u_i^{oG})/(1 - u_i^{oM}) \tag{64}$$

for the node i with the largest utilization. We then calculate u_i^{cG} for the other nodes using (15), which is justified for non-Markov models as well as Markov models, e.g., by Little's formula, (4.2.3) of Ref. 19. At least, the ratios u_i^{oG}/u_i^{oM} and $(1 - u_i^{oG})/(1 - u_i^{oM})$ can give a rough idea of how much the nonexponential service-time distributions matter.

XI. THROUGHPUT BOUNDS IN NON-MARKOV CLOSED NETWORKS

It is sometimes claimed that closed Markov models are suitable even when the service-time distributions are not exponential. In particular, it is sometimes claimed that the utilizations and throughputs, at least, do not depend critically on aspects of the service-time distributions beyond their means. Of course, this is trivially true for certain special service disciplines such as processor sharing, for which there are insensitivity results,[3-6,19,20] but with FCFS nodes the service-time distribution matters. Even with FCFS nodes, there is significant justification for this view if the service-time distributions do not depart too

drastically from the exponential distribution. However, in general, throughputs obtained with the Markov model can be very bad approximations, as we show in this section. For example, for a cyclic network with n single-server nodes, equal mean service times, and K customers, we show that the set of possible utilizations for each server is the interval $(n^{-1}, 1]$ for all $K \geq n$, whereas the utilization is $K/(n + K - 1)$ in the Markov model by (14). For large n and K and arbitrarily unfavorable service-time distributions with given means, the Markov approximation can be arbitrarily bad. The true value can be arbitrarily close to 0, while the Markov approximation is arbitrarily close to 1.

We consider the same non-Markov closed model as in Section X, containing FCFS nodes with general service-time distributions, but we restrict attention to single-server nodes. All the service times are assumed to be mutually independent and the service times at any given node are identically distributed. There is a single job class with K jobs. A job completing service at node i is routed immediately to node j with probability q_{ij}, independent of the history. The matrix $Q \equiv (q_{ij})$ is a Markov chain transition matrix, which we assume is irreducible. Consequently, there is a unique equilibrium distribution associated with Q, defined by

$$\lambda_j = \sum_{i=1}^{n} \lambda_i q_{ij}, \qquad 1 \leq j \leq n, \tag{65}$$

with $\lambda_1 + \cdots + \lambda_n = 1$. By the law of large numbers, λ_j is the long-run fraction of transitions that each customer spends in node j. The system of equations (65) is also the basic traffic-rate equations for the network of queues. The throughput or equilibrium flow rate through node i, say θ_i^e, is proportional to λ_i, i.e., $\theta_i^e = \gamma\lambda_i$ for some constant γ.

Let τ_i be the mean service time (which we assume is finite and strictly positive) and let u_i^e be the utilization (long run fraction of time that the server is busy) at node i. By Little's law, (4.2.3) of Ref. 19, or by the law of large numbers again, we know the ratio of the utilizations, i.e.,

$$u_i^e/u_j^e = \lambda_i\tau_i/\lambda_j\tau_j \tag{66}$$

for any i and j just as for the Markov model in (15).

We exhibit the infimum of the server utilizations possible for service-time distributions with the given means. As will soon be clear, the infimum is approached by quite unusual service-time distributions, so that we do not rule out the possibility that the Markov model can provide good throughput approximations for typical nonexponential service-time distributions. The idea for minimizing utilizations is really quite simple. For our model, at any time at least one server must be busy. Hence, the sum of the server utilizations must exceed unity:

$$\sum_{i=1}^{n} u_i^2 \geq 1. \tag{67}$$

A lower bound on the server utilizations is the case in which there is no concurrency, i.e., no two servers are ever busy at the same time. This lower bound is obviously valid in much greater generality. It is also attained in the case $K = 1$. It is somewhat remarkable that this lower bound is actually approached for any K with the general independent service times allowed here. This observation was apparently first made by Arthurs and Stuck.[11]

It is, in fact, not difficult to attain this lower bound asymptotically by considering special sequences of service-time distributions with a common mean that get successively more variable. In particular, for $m \geq 1$, let X_m be a random variable distributed as

$$P(X_m = m) = 1 - P(X_m = 0) = m^{-1}. \tag{68}$$

[Alternatively, $(X_m \mid X_m > 0)$ could have some other distribution with mean m, such as exponential.] Then let the service time at node i be distributed as $\tau_i X_m$.

Theorem 14: (*a*) *The infimum of the possible utilizations of server i for this closed network model over all service-times distributions with specified means is*

$$\inf u_i = \lambda_i \tau_i \Big/ \sum_{j=1}^{n} \lambda_j \tau_j.$$

(*b*) *If the service times are not all deterministic, then the infimum is not attained for* $K > 1$, *but is approached asymptotically for all nodes simultaneously as* $m \to \infty$ *using the service-time distributions of* $\tau_i X_m$ *described above.*

Proof: (a) We informally sketch the proof. For very large m, occasionally (among all service times generated) a long service time occurs at some node. With high probability, thereafter all the other customers instantaneously fly around the network until they arrive at this node, where they all wait together in queue. (There is only one server at each node.) The only other possibility, which occurs with asymptotically negligible probability as $m \to \infty$, is that one of the other customers encounters another nonzero service time before all of the customers are gathered together at the same node. This event yielding the concurrency has asymptotically negligible probability because the distribution of the number of transitions for any job to go from any node i to any other node j does not change with m. Hence, for each of the K − 1 customers there is a fixed random number of trials (with finite mean and variance) to generate a new nonzero service time, but the probability of doing so on each trial is m^{-1}. Hence, the proportion

of time during which two or more servers are simultaneously busy converges to zero as $m \to \infty$.

(b) On the other hand, it is trivial that concurrency cannot be ruled out altogether when $K > 1$ and there is some randomness. For any model with strictly positive expected service times, at least one nondeterministic distribution, and an irreducible routing matrix, concurrency occurs with positive probability. The limiting case above is not legitimate because X_m converges in distribution to the random variable X with $P(X = 0) = 1$. □

Remarks: 1. It is not necessary to have all service-time distributions be of the special form (68). It suffices to have all but one. The other one can be arbitrary. At this designated node there will be a succession of ordinary service times after which the customer usually returns immediately to the end of the queue. When a customer does get a nonzero service time elsewhere, the others get there relatively quickly with high probability when they complete service. It is again not difficult to show that the proportion of time that there is concurrency is asymptotically negligible.

2. A next step would be to obtain tighter bounds under extra conditions as, for example, in Refs. 63 through 65 and references there. However, as noted above, the special service-time distributions can be H_2 (hyperexponential: a mixture of two exponential distributions), so that does not help. It would obviously help to fix the variance though. We conjecture that the X_m distributions would yield the minimum then.

3. The infimum decreases rapidly as the number of nodes increases. The possible server utilizations are not so great with two nodes. For example, suppose that $q_{12} = q_{21} = 1$, which makes the network cyclic. Then $\lambda_1 = \lambda_2 = 1/2$ and $\inf u_1 = \tau_1/(\tau_1 + \tau_2)$. In this case the maximum is 1, which is attained with the deterministic service-time distributions for $K \geq 2$. As before, the infimum corresponds to the case $K = 1$. In the case of balanced loads, the utilization of each server must lie between one-half and one. It is useful to recall that this two-server cyclic network is equivalent to an M/G/1/K−1 queueing model when one of the two service-time distributions is exponential (see Ref. 66, p. 33 of Ref. 67, and Ref. 2). The M/G/1/K−1 model means that we have an external Poisson arrival process, a single queue with one server and an additional waiting room of size $K - 1$. Since we approach the infimum if all but one service-time distribution is of the special kind, the infimum is also valid for M/G/1/K−1 queueing model.

XII. MORE NUMERICAL COMPARISONS

We have indicated that the approximation methods should perform better for larger closed networks, but it is nevertheless useful to

compare their performance for smaller ones. It is certainly important to realize the limitations of these procedures. They often perform poorly for small networks.

It is particularly convenient to consider two-node closed networks because these networks are equivalent to special single-node models that have been studied extensively and for which there are tables of exact values.

12.1 Two single-server nodes

As we noted in Section XI, the closed model with K jobs (all of one class) and two single-server nodes, one of which has an exponential service-time distribution, is equivalent to an M/G/1 model with a finite waiting room of size $K-1$. Similarly, the two-node closed model with K jobs and one IS node having exponential service-time distributions is equivalent to finite-source M/G/1 model with K sources. Tables for M/G/1 models with finite waiting room and finite sources are contained in Ref. 67, for example.

Table V here displays the exact values and various approximations for the throughput and the expected equilibrium number of jobs present in an M/G/1 model having a waiting room of size 10. This

Table V—A comparison of exact throughput and mean number in system in the M/G/1/10 model having a finite waiting room of size 10 with approximations based on the bottleneck method and the FPM method for $C = M$, D, and H_2

Arrival Rate	Exact Values		Bottleneck Method		FPM Method		Predicted θ With (64)
	θ^e	EN_1	θ^o	EN_1	θ^o	EN_1	
(a) Exponential Service Times (M)							
0.50	0.4999	1.00	0.500	1.00	0.455	0.84	
0.75	0.7418	2.61	0.750	3.00	0.674	2.07	
1.00	0.9167	5.50	1.000	∞	0.846	5.50	
1.40	0.9928	8.72	1.000	8.50	0.902	9.19	
2.00	0.9998	10.00	1.000	10.00	0.910	10.16	
(b) Deterministic Service Times (D)							
0.50	0.5000	0.75	0.500	0.75	0.461	0.65	0.500
0.75	0.7494	1.85	0.750	1.88	0.695	1.43	0.744
1.00	0.9538	5.57	1.000	∞	0.912	4.93	0.952
1.40	0.9998	9.61	1.000	9.39 (9.69)	0.973	9.68	0.998
2.00	1.0000	10.37	1.000	10.25 (10.61)	0.987	10.38	1.000
(c) Hyperexponential Service Times With $c_s^2 = 2.25$ and Balanced Means							
0.50	0.4983	1.27	0.500	1.31	0.449	1.05	0.500
0.75	0.7248	3.01	0.750	4.41	0.651	2.71	0.739
1.00	0.8829	5.34	1.000	∞	0.787	6.00	0.885
1.40	0.9791	8.15	1.000	7.38	0.833	9.06	0.987
2.00	0.9985	9.75	1.000	9.69	0.840	10.10	1.000

corresponds to a closed network with a population of 11 and two single-server nodes. Three service-time distributions are considered: exponential, deterministic, and hyperexponential (mixture of two exponentials). The hyperexponential distribution has squared coefficient of variation $c_s^2 = 2.25$ and balanced means (see p. 8 of Ref. 67 or Section 3 of Ref. 68). The service time is set equal to 1 and five arrival rates are considered: 0.50, 0.75, 1.00, 1.40, and 2.00. The arrival rate of 2.00 (1.40) corresponds to a traffic intensity of 0.50 (0.71) when the nodes are switched. In each case, the FPM and bottleneck approximations are displayed in addition to the exact values. For the nonexponential service times, the refinement in (64) is also displayed.

The exact values come from Tables 5.1.6, 5.2.6, and 5.4.12 in Section II.5 of Ref. 67, using the FIFO or FCFS discipline. The approximations are obtained using the GI/G/1 formulas (47) and (44) in Ref. 5 with g in (45) of Ref. 5 set equal to 1. The approximate values using the Krämer-and-Langenbach-Belz correction term in (45) of Ref. 7 are given in parentheses to the right of the other values in Table V (b).

There are several important conclusions to draw from Table V. First, as we should expect from Section III, the FPM method performs poorly, much worse than the bottleneck method. However, it is important to remember that this small network tends to be a worst case for the FPM method. It is also significant that the refinement suggested in (64) produces quite accurate results. With this job population (waiting room size), the bottleneck method seems to work reasonably well as long as the utilization of the bottleneck node is no more than about 0.75.

It is also useful, to consider the Finite-Writing-Room (FWR) refinement introduced in Section 1.3 in the context of Table Va. When combined with the bottleneck procedure, the FWR refinement obviously makes the approximation exact when $n = 2$. The FPM method with the FWR refinement is also exact in the special case of equal service rates. When $n = 2$, we must have $E\bar{N}_i^o = EN_i^o = K/2$ by the FPM method. With the FWR method, this implies that $\rho_i = 1$ and $P(\bar{N}_i^o = K) = 1/(K+1)$, so that $\bar{n}_i^o = u_i^c = K/(K+1) > K/(K+2) = u_i^o$, using (11) and (14). However, more generally the FPM/FWR method does not perform well, at least for the mean numbers at each node, when $n = 2$. For example, when the arrival rate is 0.50, the FPM/FWR approximations are $\theta^o = 0.69$ and $EN_1 = 2.17$. However, applying (3), we obtain $\overline{\theta}^* = 0.495$ as a lower bound on the throughput.

From Table Vb we see that the bottleneck method continues to perform well for the deterministic service-time distribution. In fact, the bottleneck approximation is clearly much better than using the exact M/M/1 values in Table Va as an approximation, which is often what is done in practice.

However, from Table Vc we see that the quality of the bottleneck approximation deteriorates when we consider the more variable hyperexponential service-time distribution. Of course, the throughputs are always close and the mean queue lengths are good when the traffic intensity at the bottleneck node is 0.5, but when the traffic intensity is 0.70 or 0.75, the open-network view exaggerates the impact of the greater variability. The fixed population in the closed model tends to damp the effect.

We can also see what happens as we change the service-time variability in Table V. From the exact values, we see that the throughput decreases in every case, but the throughputs are never near the lower bound in Theorem 14. We also see that the expected number of jobs at that node decreases when the arrival rate is greater than 1.00. This phenomenon was observed by Bondi[61] and is further discussed in Bondi and Whitt.[62] Briefly, the explanation is that under moderate to heavy loads increased variability in the service-time distribution often has a greater impact on other nodes via their arrival processes than on the congestion at the given node. It is significant that this qualitative behavior is captured by both the bottleneck and FPM methods using QNA. However, the FPM method fails to capture this bottleneck phenomenon in other cases. As noted in Refs. 61 and 62, the bottleneck phenomenon is useful to test procedures for approximately solving non-Markov closed networks.

Approximately characterizing the variability of arrival processes in a tightly coupled closed network such as the two-node model being discussed is difficult because of the constraint on the total population. If the utilization of a server is high, then the interdeparture times are distributed approximately the same as the service times, but the population constraint tends to induce negative correlations among the interdeparture times: several long (short) times are more likely to be followed by a short (long) one. Hence, the effective variability of an arrival process, e.g., as described by the asymptotic method in Ref. 68, is likely to be considerably less in a closed network than in an open one. This is the reason for developing heuristic procedures to reduce the variability parameters of the arrival processes in the approximation method.

12.2 One single-server node and one infinite-server node

Table VI displays exact and approximate results for a two-node network containing an IS node with exponential service-time distributions. In this case, the service-time distribution at the single-server node is always exponential. It is easy to apply the FPM approximation to other cases, but we had no convenient tables. The exact values from Table 2.10.7 of Ref. 67 are obtained by specifying the population

Table VI—A comparison of exact throughput and mean number in system in the M/G/1 queue having a finite source of size 10 with approximations based on the FPM method

	Exact Values	FPM	First Upper Bound
(a) Total Population 10 and Utilization 0.50			
Arrival rate per idle source	0.0547	0.0547 given	0.0547 given
Throughput or utilization	0.500 given	0.494	0.547
Expected number in system, EN_1	0.86	0.98	1.21
(b) Total Population 10 and Utilization 0.75			
Arrival rate per idle source	0.0927	0.0927 given	0.0927 given
Throughput or utilization	0.750 given	0.705	0.927
Expected number in system, EN_1	1.91	2.39	12.70

(number of sources) and the throughput. Paralleling Table V, we consider two cases: a population of 10 and throughputs of 0.50 and 0.75. As in Table V the service time is set equal to 1. The population and throughput determines the arrival rate per idle source (individual service rate at the IS node). This is the starting point for the FPM approximation, which is obtained from (28) or (38). The conditions are clearly much less favorable in Table VI than in Tables I and IV; the ratio of service rates (IS/other) was much less before. Nevertheless, the FPM method works quite well, at least in the case of utilization 0.50. From Tables V and VI, we see that the FPM method does indeed perform better with the IS node. Related numerical comparisons are contained in Ref. 69. The overall performance here based on the FPM method is somewhat better than the performance in Ref. 69, which is based on matching the server utilization. The performance in Tables I and IV is much better than we might at first expect from Ref. 69, but recall that in Tables I and IV, as the total population decreases the server utilization decreases because the service rate at the IS node is held fixed. Nevertheless, the tables in Ref. 69 help assess how well the FPM method will perform for a small network with an IS node.

XIII. CONCLUSIONS

In this paper we identified and investigated three situations in which open queueing network models should provide good approximations for more difficult closed queueing network models:

1. When the closed network has many nodes (Sections II through V, X),

2. When the closed network contains a "decoupling" infinite-server (IS) node with a relatively low service rate (see Sections VI through VIII),

3. When the closed network contains a non-IS bottleneck node under a fairly heavy load (Section IX).

The suggested approximation procedures in these situations are not the same, however. In Case 3 we remove the bottleneck server and replace its departure process by an external arrival process, which is determined solely by the number of servers and the service-time distribution at the bottleneck queue. The arrival rate is the maximum possible service rate from the bottleneck node; we do not use the FPM method. In contrast, in Case 1 no nodes are removed from the closed network. As described in Section X, an entry node is selected and the external arrival process there depends on the entire network in a rather complicated way. The arrival rate is determined by the FPM method. The variability parameter of the arrival process, using QNA,[7] is chosen so that the variability parameter of the external arrival process agrees with the variability parameter of the departure process from the network.

It is interesting that the suggested procedure for Case 2 can be regarded as a variation of either the procedure for Case 1 or the procedure for Case 3. On the one hand, the procedure for Case 1 can be applied without change to Case 2. As described in Section VI, the suggested procedure coincides with the FPM method. However, in Case 2 we know that the departure process from the IS node is approximately a Poisson process. Hence, it is natural to implement the FPM method for Case 2 by replacing the departure process from the decoupling IS node by an external Poisson arrival process. We then use the FPM method to determine the appropriate external arrival rate, but we do not have to worry about the variability parameter; we just set it equal to 1. If instead we used the FPM method as described for Case 1 and we selected an entry node for an external arrival process, then we would need to specify the variability parameter of the external arrival process there. In general, in Case 2 the arrival processes to other nodes need not be approximately Poisson. However, if we apply the standard FPM method for Case 1 to Case 2, then the results should be very similar because in Case 2 the FPM method will make the variability parameter of the departure process from the decoupling IS node nearly 1.

We can also think of the procedure for Case 2 as a modification of the procedure for Case 3. The decoupling IS node also acts as a bottleneck queue. Hence, as described in Section VI, we can analyze Case 2 by removing the IS node from the closed network and replacing its department process by an external arrival process. Because of the

nature of this particular bottleneck queue, i.e., because there are many servers each with low service rate, it is appropriate to make the external arrival process a Poisson process. Incidentally, we would do this in Case 3 too if there were many servers, but finitely many, each with low service rate.

If we apply the procedure for Case 3 directly to Case 2, then we let the arrival rate of the external Poisson process be the maximal possible service rate from the IS node, which corresponds to the first upper bound described in Section VI. The suggested modification is to let the arrival rate of the external Poisson arrival be such that this arrival rate would equal the departure rate at the IS node if it were included in the network. As indicated in Section VI, this modification turns out to coincide with the FPM method.

It is important to recognize that the three situations above do not nearly cover all possibilities. As indicated in Section I, in some cases an open model might also be reasonable from direct modeling considerations; often the closed model is not entirely appropriate. However, it is clear from the analysis and examples here that open models do not always produce reasonable, let alone good, approximations for closed models. For closed networks with few nodes, few servers per node and few jobs, the open-model approximations for closed models here tend to perform poorly. Further experimentation is needed to better understand the appropriate regions for each procedure. As with any approximation tool, it is very helpful in applications to make a few initial benchmark comparisons with simulations to determine the actual quality of the approximations in that context.

The specific models discussed in this paper have been relatively elementary. Many of the theorems only relate open and closed Markov Jackson networks. The major complexity considered was nonexponential FCFS servers and the associated network model treated by QNA. It is important to realize, however, that the ideas apply much more broadly. As discussed by Zahorjan,[22] these open-model approximations for closed models can be used as modules or subroutines in more complicated approximation procedures, e.g., based on network decomposition. As illustrated by Fredericks,[40] the ideas also apply directly to closed models with other complicating features such as priorities.

XIV. ACKNOWLEDGMENTS

I am grateful to A. T. Seery for writing the QNA software and to A. B. Bondi, A. A. Fredericks, H. Heffes, M. Segal, and D. R. Smith for their interest and suggestions.

REFERENCES

1. J. R. Jackson, "Jobshop-like Queueing Systems," Manage. Sci., *10*, No. 1 (October 1963), pp. 131–42.

2. W. J. Gordon and G. F. Newell, "Closed Queueing Systems With Exponential Servers," Oper. Res., *15*, No. 2 (March-April 1967), pp. 254–65.
3. F. Baskett, M. Chandy, R. Muntz, and J. Palacios, "Open, Closed and Mixed Networks of Queues With Different Classes of Customers," J.A.C.M., *22*, No. 2 (April 1975), pp. 248–60.
4. L. Kleinrock, *Queueing Systems, Volume 2: Computer Applications*, New York: Wiley Interscience, 1976.
5. F. P. Kelly, *Reversibility and Stochastic Networks*, New York: Wiley, 1979.
6. C. H. Sauer and K. M. Chandy, *Computer Systems Performance Modeling*, Englewood Cliffs, NJ: Prentice-Hall, 1981.
7. W. Whitt, "The Queueing Network Analyzer," B.S.T.J., *62*, No. 9 (November 1983), pp. 2779–815.
8. K. G. Ramakrishnan and D. Mitra, "An Overview of PANACEA, A Software Package for Analyzing Markovian Queueing Networks," B.S.T.J., *61*, No. 10 (December 1982), pp. 2849–72.
9. J. Zahorjan et al., "Balanced Job Bound Analysis of Queueing Networks," Commun. A.C.M., *25*, No. 2 (February 1982), pp. 134–41.
10. D. L. Eager and K. C. Sevcik, "Performance Bound Hierarchies for Queueing Networks," A.C.M. Trans. Comput. Syst., *1*, No. 2 (May 1983), pp. 99–115.
11. E. Arthurs and B. W. Stuck, "Upper and Lower Bounds on Mean Throughput Rate and Mean Delay in Memory-Constrained Queueing Networks," B.S.T.J., *62*, No. 2 (February 1983), pp. 541–81.
12. F. Spitzer, *Principles of Random Walk*, Princeton, NJ: Van Nostrand, 1964.
13. J. McKenna, D. Mitra, and K. Ramakrishnan, "A Class of Closed Markovian Queueing Networks: Integral Representation, Asymptotic Expansions, Generalizations," B.S.T.J., *60*, No. 5 (May-June 1981), pp. 599–641.
14. J. McKenna and D. Mitra, "Integral Representations and Asymptotic Expansions for Closed Markovian Queueing Networks: Normal Usage," B.S.T.J., *61*, No. 5 (May-June 1982), pp. 661–83.
15. D. J. Jagerman, "Nonstationary Blocking in Telephone Traffic," B.S.T.J., *54*, No. 3 (March 1975), pp. 625–61.
16. W. Whitt, "Heavy-Traffic Approximations for Service Systems With Blocking," AT&T Bell Lab. Tech. J., *63*, No. 5 (May 1984), pp. 689–708.
17. D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, New York: Wiley, 1974.
18. W. Whitt, unpublished work.
19. P. Franken, D. König, U. Arndt, and V. Schmidt, *Queues and Point Processes*, Berlin: Akademie-Verlag, 1981.
20. D. Y. Burman, "Insensitivity in Queueing Systems," Advance. Appl. Probab., *13*, No. 4 (December 1981), pp. 846–59.
21. Information Research Associates, "User's Manual for CADS," Austin, TX, 1978.
22. J. Zahorjan, "Workload Representations in Queueing Models of Computer Systems," *Proc. ACM Sigmetrics Conf.*, Minneapolis, August 1983, pp. 70–81.
23. J. Keilson, *Markov Chain Models—Rarity and Exponentiality*, New York: Springer-Verlag, 1979.
24. D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, New York: Wiley, 1983.
25. S. Karlin and Y. Rinott, "Classes of Orderings of Measures and Related Correlation Inequalities. II. Multivariate Reverse Rule Distributions," J. Multivar. Anal., *10*, No. 4 (December 1980), pp. 499–516.
26. H. W. Block, T. H. Savits, and M. Shaked, "Some Concepts of Negative Dependence," Ann. Probab., *10*, No. 3 (August 1982), pp. 765–72.
27. K. Joag-Dev and F. Proschan, "Negative Association of Random Variables, With Applications," Ann. Statist., *11*, No. 1 (March 1983), pp. 286–95.
28. P. Billingsley, *Convergence of Probability Measures*, New York: Wiley, 1968.
29. W. Whitt, "Some Useful Functions for Functional Limit Theorems," Math. Oper. Res., *5*, No. 1 (February 1980), pp. 67–85.
30. T. C. Brown, "Some Poisson Approximations Using Compensators," Ann. Probab., *11*, No. 3 (August 1983), pp. 726–44.
31. W. Whitt, "Continuity of Generalized Semi-Markov Processes," Math. Oper. Res., *5*, No. 4 (November 1980), pp. 494–501.
32. W. Whitt, "The Continuity of Queues," Advance. Appl. Probab., *6*, No. 1 (March 1974), pp. 175–83.
33. T. Kamae, U. Krengel, and G. L. O'Brien, "Stochastic Inequalities on Partially Ordered Spaces," Ann. Probab., *5*, No. 6 (December 1977), pp. 899–912.
34. W. Whitt, "Comparing Counting Processes and Queues," Advance. Appl. Probab.,

13, No. 1 (March 1981), pp. 207–20.
35. P. Whittle, "Equilibrium Distributions for an Open Migration Process," J. Appl. Probab., *5,* No. 3 (December 1968), pp. 567–71.
36. T. C. Brown and P. K. Pollett, "Some Distributional Approximations in Markovian Queueing Networks," Advance. Appl. Probab., *14,* No. 3 (September 1982), pp. 654–71.
37. B. Hajek, "The Proof of a Folk Theorem on Queueing Delay With Applications to Routing in Networks," J.A.C.M. *30,* No. 4 (October 1983), pp. 834–51.
38. W. A. Massey, "An Operator Analytic Approach to the Jackson Network," J. Appl. Probab., *21,* No. 2 (June 1984), pp. 379–93.
39. J. B. Goodman and W. A. Massey, unpublished work.
40. A. A. Fredericks, unpublished work.
41. W. Feller, *An Introduction to Probability Theory and Its Applications,* Vol. I, Third Ed., New York: Wiley, 1968.
42. S. Halfin and W. Whitt, "Heavy-Traffic Limits for Queues With Many Exponential Servers," Oper. Res., *29,* No. 3 (May-June 1981), pp. 567–88.
43. J. W. Cohen, *The Single Server Queue,* Amsterdam: North-Holland, 1969.
44. D. Stoyan and H. Stoyan, "Monotonieeigenschaften der Kundenwartezeiten im Model GI/G/1," Z. Angew. Math., *49,* No. 12 (1969), pp. 729–34.
45. T. Rolski and D. Stoyan, "On the Comparison of Waiting Times in GI/G/1 Queues," Oper. Res., *24,* No. 1 (January-February 1976), pp. 197–200.
46. J. Keilson, "A Threshold for Log-Concavity for Probability Generating Functions and Associated Moment Inequalities," Ann. Math. Statist. *43,* No. 5 (September 1972), pp. 1702–8.
47. W. Whitt, "Uniform Conditional Stochastic Order," J. Appl. Probab., *17,* No. 1 (March 1980), pp. 112–23.
48. W. Whitt, "Multivariate Monotone Likelihood Ratio and Uniform Conditional Stochastic Order," J. Appl. Probab., *19,* No. 3 (September 1982), pp. 695–701.
49. W. Whitt, unpublished work.
50. S. Karlin and Y. Rinott, "Classes of Orderings of Measures and Related Correlation Inequalities. I. Multivariate Total Positivity," J. Multivar. Anal., *10,* No. 4 (December 1980), pp. 476–98.
51. E. Cinlar, "Superposition of Point Processes," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications,* P. A. W. Lewis, ed., New York: Wiley, 1972, pp. 549–606.
52. F. Böker and R. Serfozo, "Ordered Thinnings of Point Processes and Random Measures," Stoch. Proc. Appl., *15,* No. 2 (July 1983), pp. 113–32.
53. D. Sonderman, "Comparing Semi-Markov Processes," Math. Oper. Res., *5,* No. 1 (February 1980), pp. 110–9.
54. P. J. Schweitzer, "Bottleneck Determination in Networks of Queues," *Applied Probability—Computer Science: The Interface,* Vol. I, R. L. Disney and T. J. Ott, eds., Boston: Brikhäuser, pp. 471–85.
55. M. I. Reiman, "Open Queueing Networks in Heavy Traffic," Math. Oper. Res., *9,* No. 3 (August 1984), pp. 441–58.
56. M. I. Reiman, "Some Diffusion Approximation With State Space Collapse," *Proc. Int. Seminar on Modeling and Performance Evaluation Methodology,* New York: Springer-Verlag, 1983.
57. M. Reiser and H. Kobayashi, "Accuracy of the Diffusion Approximation for Some Queueing Systems," IBM J. Res. Develop., *18* (March 1974), pp. 110–24.
58. A. W. Shum and J. Buzen, "A Method for Obtaining Approximate Solutions to Closed Queueing Networks With General Service Times," in *Modeling and Performance Evaluation of Computer Systems,* H. Beilner and E. Gelenbe, eds., Amsterdam: North-Holland, 1978.
59. A. W. Shum, *Queueing Models for Computer Systems with General Service Time Distributions,* New York: Garland, 1980.
60. S. K. Tripathi, "On Approximate Solution Techniques for Queueing Network Models of Computer Systems," Ph.D. dissertation, Department of Computer Sciences, University of Toronto, 1981.
61. A. B. Bondi, "Incorporating Open Queueing Models Into Closed Queueing Network Algorithms," Ph.D. dissertation, Department of Computer Sciences, Purdue University, 1984.
62. A. B. Bondi and W. Whitt, unpublished work.
63. W. Whitt, "On Approximations for Queues, I: Extremal Distributions," AT&T Bell Lab. Tech. J., *63,* No. 1 (January 1984), pp. 115–38.
64. J. G. Klincewicz and W. Whitt, "On Approximations for Queues, II: Shape Con-

straints," AT&T Bell Lab. Tech. J., *63*, No. 1 (January 1984), pp. 139–61.
65. W. Whitt, "On Approximations for Queues, III: Mixtures of Exponential Distributions," AT&T Bell Lab. Tech. J., *63*, No. 1 (January 1984), pp. 163–75.
66. M. Reiser and H. Kobayashi, "The Effects of Service Time Distributions on System Performance," *Information Processing 74*, Amsterdam: North-Holland, 1974, pp. 230–4.
67. P. Kühn, *Tables on Delay Systems*, Stuttgart: Institute of Switching and Data Technics, University of Stuttgart, 1976.
68. W. Whitt, "Approximating a Point Process by a Renewal Process, I: Two Basic Methods," Oper. Res., *30*, No. 1 (January-February 1982), pp. 125–47.
69. J. P. Buzen and P. S. Goldberg, "Guidelines for the Use of Infinite Source Queueing Models in the Analysis of Computer System Performance," Proc. AFIPS Nat. Comput. Conf., *43*, 1974, pp. 371–4.

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. His work focuses on stochastic processes and congestion models.