The Influence of Service-Time Variability in a Closed Network of Queues

André B. Bondi *
Department of Computer Science, University of California at Santa Barbara, Santa Barbara, CA 39106, U.S.A.

Ward Whitt
AT&T Bell Laboratories, Holmdel, NJ 07733, U.S.A.

Received 12 July 1984
Revised 11 June 1985

This paper describes the effect of service-time variability on the standard performance measures of a closed network of single-server queues with the first-come first-served discipline and one job class. Several service-time variability principles are proposed to serve as rough practical guidelines. The most interesting one states that the mean queue length at a bottleneck queue typically decreases when the variability of the service time at that queue is increased. The principles are supported here by numerical examples and theorems in special cases. The principles are also applied to test approximation procedures.

Keywords: Queues, Queueing Networks, Nonexponential Service Times, Approximations.

André B. Bondi was born in Forest Hills, New York, in 1955. He is currently an Assistant Professor in the Department of Computer Science at the University of California, Santa Barbara. He received a B.Sc. in Mathematics from the University of Exeter in 1976, an M.Sc. in Statistics from University College London in 1977, and M.S. and Ph.D. degrees in Computer Science from Purdue University in 1981 and 1984, respectively. From 1977 and 1979, he worked as a consultant in performance evaluation at Lociga Ltd., London. Mr. Bondi has spent summers at BGS Systems, the IBM Thomas, J. Watson Research Center, and Pritsker and Associates. His research interests include queueing theory and performance evaluation of computer communication networks. He is a member of ACM (including SIGMETRICS, SIGCOMM and SIGOPS) and a Fellow of the Royal Statistical Society.

* This work was done while the author was completing his Ph.D. Dissertation at the Department of Computer Sciences at Purdue University, U.S.A.

North-Holland
Performance Evaluation 6 (1986) 219–234

1. Introduction and summary

This paper discusses the relationship between service-time variability and the standard performance measures in a closed network of single-server queues with the FCFS (first-come-first-served) discipline. The model is the standard Jackson [9] or Gordon/Newell [8] closed Markovian single-class queueing network, modified by having only single-server queues and by allowing nonexponential service-time distributions. For this model, we propose some service-time variability principles. These variability principles are intended to serve as rough practical guidelines. They represent general tendencies observed in numerical examples, that are supported by theorems in special cases.

We also discuss the implication of the service-time variability principles for procedures to approximately analyze closed networks of single-server FCFS queues having nonexponential service-time distributions. Obviously, it is desirable for approximation procedures to be consistent with these variability principles. From this perspective, we investigate several specific approximation procedures: the Reiser mean-value-analysis (MVA) procedure [19], the Chandy/Herzog/Woo (CHWS) device-complement procedure [5], the Shum/Buzen extended-product-form (EPF) procedure [27,28], and the Marie device-complement procedure [15,16].

To set the stage for our investigation of closed networks, and for its own sake, we first discuss the

Ward Whitt was born in Buffalo, New York, in 1942. He has been a member of the technical staff at AT&T Bell Laboratories since 1977. He received an A.B. in Mathematics from Dartmouth College in 1964 and a Ph.D. in Operations Research from Cornell University in 1969. He taught at Yale University from 1969 to 1977. His research has focused primarily on queues.

somewhat clearer situation of open networks. It is generally understood that the congestion in an open network of single-server queues with unlimited waiting space and the FCFS discipline will increase if any service-time distribution becomes more variable (more spread out or dispersed). We obtain a concrete expression of this idea if we measure congestion at a queue by the mean (expected equilibrium) queue length and if we measure variability of the service-time distribution by its CV (coefficient of variation, defined as the standard deviation divided by the mean). Of course, the variability of a probability distribution with fixed mean and the impact of that distribution as a service-time distribution on the behavior of a queue are not completely determined by the CV [3,14], but the CV is certainly a major factor which is useful for practical purposes [11,34]. The following are familiar service-time variability principles for open networks.

Service-time variability principles for open networks

(OS1) *Each mean queue length is a nondecreasing function of the service-time* CV *at that queue.*

(OS2) *Each mean queue length is also a nondecreasing function of the service-time* CVs *at all other queues.*

Variability principle (OS2) is somewhat less apparent than (OS1). The reason for (OS2) is that increased variability of a service-time distribution typically causes increased variability of the departure process from that queue, which in turn causes increased variability in the arrival processes at the other queues that can be reached from the given queue. Thus, principle (OS2) is a consequence of the following arrival-process variability principle.

Arrival-process variability principle for open networks

(OA1) *The mean queue lengths are nondecreasing functions of each arrival-process* CV.

Of course, these variability principles are not valid as theorems without qualifications. Throughout this paper we assume that there is a single job class and that a job completing service at queue i goes next to queue j with probability q_{ij}, independent of the history of the system. As usual, we assume that the service times at each queue and interarrival times of each external arrival process (if any) are mutually independent. Moreover, the service times at each queue have a common distribution and the interarrival times of each external arrival process have a common distribution. However, even with these additional specifications, the variability principles are not valid without qualifications, but they are usually appropriate for practical purposes. We discuss them further in Section 2.

In the associated closed network, there are no external arrivals; instead, a fixed population of jobs circulates around the network. The two service-time variability principles above cannot be extended to closed networks because the total population is fixed. Indeed, we should not expect that principle (OS2) would extend to closed networks, and it does not. It may at first seem surprising, but even principle (OS1) does not extend to closed networks. Of course, principle (OS1) does often apply to closed networks, but there is a systematic exception. Increasing the service-time variability at a bottleneck queue typically causes the mean queue length there to decrease, when the population is not too small. The bottleneck queue is the queue with the largest utilization, where the utilization is the long-run fraction of time (or probability) that the server is busy. This 'bottleneck phenomenon' was observed and verified by simulations, analytic approximations and exact analytic methods in [2]. We present some of the evidence here. We also propose four service-time variability principles for closed networks.

Service-time variability principles for closed networks

(CS1) *The mean queue lengths tend to be relatively insensitive to changes in the service-time* CVs, *provided that the network population is not too large.*

(CS2) *The mean queue length at a nonbottleneck queue is a nondecreasing function of the service-time* CV *at that queue.*

(CS3) *The mean queue length at a bottleneck queue is a nonincreasing function of each service-time* CV, *provided that the number of jobs in the network is sufficiently*

large; otherwise, it tends to be a nondecreasing function.

(CS4) *The vector of server utilizations is a nonincreasing function of the service-time* CV *at any queue.*

Remarks. (1) Principle (CS1) is somewhat vague. One way to be more precise is to compare with related open models: Typically, the impact on the mean queue lengths of the same change in service-time CV is much less in a closed model. However, the influence increases as the population increases.

(2) The fact that the mean queue length need not increase when the service-time variability is increased at that node has no doubt been observed several times before. For example, this phenomenon was observed for the finite-capacity M/G/1 queue by Lavenberg in [13, Table 1 and p. 506]. Principle (CS3) and later discussion here help explain this phenomenon.

(3) Since the ratio of the utilizations at different queues is determined by the routing probabilities and the service rates, even with nonexponential service-time distributions, all the utilizations will respond the same way; there is only one degree of freedom for principle (CS4).

Principles (CS1), (CS2), and (CS4) are as expected, but principle (CS3) seems to be an anomaly. It can be explained as follows. Increasing the service-time variability at one queue tends to both increase the mean queue length there and increase the variability of the departure process. However, in a network of queues, the departure process from one queue contributes to the arrival processes at other queues. Since increased variability of the arrival process also tends to increase the mean queue length, as in (OA1), increasing the service-time variability at one queue tends to increase the mean queue length at all queues that can be reached from that queue. However, this cannot happen in a closed network without drawing jobs away from the queue in question, since the number of jobs in the network is constant. In a closed network the outcome depends on whether the increased service-time variability has a greater impact at the given queue or at all other queues. The impact on other queues is most pronounced when service-time variability is increased at a heavily used node. In particular, if the server in question is the bottleneck in a closed network with a fairly large population, then it turns out that the effect of increased service-time variability is usually to draw jobs away from the bottleneck queue.

The remainder of this paper is organized as follows. Section 2 discusses the variability principles for open networks. Section 3 discusses numerical evidence in support of the service-time variability principles for closed networks. Section 4 indicates how the service-time variability principles for closed networks can be explained, in part, by relating a closed network with a large population to the associated open network obtained by replacing the bottleneck queue by an external source. Section 4 also contains a proof of principles (CS2) and (CS3) in the special case of a cyclic two-queue closed network. Section 5 contains a proof of principle (CS4) in the special case of a cyclic two-queue network in which the bottleneck service-time distribution is exponential. Finally, Section 6 discusses the implications of the service-time variability principles for procedures to approximately analyze closed networks.

2. The variability principles for open networks

In this section, we give partial support for the variability principles for open networks, first by considering the special case of a single queue. Principle (OS1) is easily verified in the special case of a single M/G/1 queue (with Poisson arrival process) because the mean queue length has a simple formula which depends on the service-time distribution only through its first two moments. For the GI/G/1 queue with renewal arrival process, principle (OS1) is good, roughly speaking, but it is obviously false without extra conditions because, for any given interarrival-time distribution and first two moments of the service time, there is a set of possible service-time distributions having those moments and a set of associated mean queue lengths. However, it is possible to show that the set of possible mean queue lengths is indeed a nondecreasing function of the service-time CV. We interpret $A_1 \leqslant A_2$ for sets of real numbers as $\inf A_1 \leqslant \inf A_2$ and $\sup A_1 \leqslant \sup A_2$. For more on this perspective involving sets of possible operating characteristics, see [34].

Theorem 2.1. *For the* GI/G/1 *queueing model with any fixed interarrival-time distribution and any fixed mean service-time, the set of possible expected equilibrium queue lengths is nondecreasing in the service-time* CV.

Theorem 2.1 is a rather direct consequence of a basic stochastic comparison result for the GI/G/1 queue [29] that provides additional support for the variability principle for open networks (see also [23,30]). One random variable X_1 or its cdf F_1 is said to be *less variable* than another random variable X_2 or its cdf F_2, denoted by $F_1 \leq_v F_2$ or $X_1 \leq_v X_2$, if $E\,g(X_1) \leq E\,g(X_2)$ for all convex real-valued functions g for which the expectations are defined. Since $g(x) = x$ and $g(x) = -x$ are both convex, $E(X_1) = E(X_2)$ and $\mathrm{CV}(X_1) \leq \mathrm{CV}(X_2)$ if $X_1 \leq_v X_2$. However, the converse is not valid: $\mathrm{CV}(X_1) \leq \mathrm{CV}(X_2)$ and $E\,X_1 = E\,X_2$ do not imply that $X_1 \leq_v X_2$. For two GI/G/1 queues with common interarrival-time distribution, the mean queue lengths are indeed ordered if the service-time distributions are ordered by $\leq_v$. In fact, the equilibrium queue length distributions are ordered in the sense of expectations of all nondecreasing convex functions. Furthermore, this result has been extended to several single-server queues in series [17].

Proof of Theorem 2.1. Given any service-time cdf F_1 with positive mean m_1 and arbitrary CV_1, it is possible to construct another service-time cdf F_2 such that $F_1 \leq_v F_2$ (so that $m_1 = m_2$) and $\mathrm{CV}_1 < \mathrm{CV}_2$ by a simple transfer of mass. (Use the property that $X_1 \leq_v X_2$ if and only if X_2 is distributed as $X_1 + Y$ where $E(Y \mid X_1) = 0$ with probability one.) Moreover, we can make CV_2 assume any value. Similarly, if $\mathrm{CV}_1 > 0$, it is possible to construct F_2 so that $F_1 \geq_v F_2$ with $\mathrm{CV}_1 > \mathrm{CV}_2$. To carry out the proof, choose a sequence of service-time cdf's $\{F_n\}$ with fixed m_1 and CV_1 such that the associated mean queue lengths approach the supremum possible given m_1 and CV_1. Then construct an associated sequence of cdf's $\{G_n\}$ with $\mathrm{CV}(G_n) = \mathrm{CV}_2 > \mathrm{CV}_1$ and $F_n \leq_v G_n$ for all n. As a consequence of the Stoyan [29] ordering, the mean queue lengths associated with G_n are larger than the mean queue lengths associated with F_n. Hence, the supremum of the mean queue lengths given m_1 and CV_1 is less than or equal to the supremum given m_1 and CV_2. A similar argument applies to the infimum. □

Remark. Principle (OS1) does not extend to multi-server queues. Counter-examples for GI/G/s queues with $s > 1$ appear in [38,22]. We conjecture that (OS1) is valid for M/G/s queues, but it has not yet been proved. It is valid for several approximation procedures for M/G/s queues (see [35] and references cited therein). More generally, we conjecture that the principle extends when the arrival processes tend to be no more variable than the Poisson process. Roughly speaking, this means when the interarrival-time CV is less than or equal to 1.

Since the basic stochastic comparison result for GI/G/1 queues in [29] applies to interarrival times as well as service-times, there is a corresponding theorem in support of principle (OA1), proved in the same way.

Theorem 2.2. *For the* GI/G/1 *queueing model with any fixed service-time distribution and any fixed mean interarrival time, the set of possible expected equilibrium queue lengths in nondecreasing in the interarrival-time* CV.

We *conjecture* that Theorems 2.1 and 2.2 extend to any 'GI/G/1 network' of the kind we are considering, i.e., any open network of single-server FCFS queues with one job class, unlimited waiting rooms and general service-time distributions, provided that the service times are independent and identically distributed at each queue, the external arrival processes are independent renewal processes, the service times at different queues are mutually independent and independent of the arrival processes, and there are independent routing probabilities. Even if the conjecture turns out not to be correct as a theorem, we believe that it describes the typical situation.

In fact, all numerical evidence known to us supports the variability principles for open networks of single-server FCFS queues, with the proper qualifications. As a consequence, approximation procedures for open networks that are based only on means and CVs invariably are consistent with the variability principles above. This is certainly true of the Queueing Network Analyzer (QNA) described in [33] and all previous algorithms mentioned therein. For QNA, it is easy to show that the vector of mean queue lengths is an nondecreasing function of the vector of

service-time CVs and external arrival-process CVs. Moreover, the approximation formulas there make it easy to quantify the influence of variability. Approximation formulas (33), (36) and (38) in [33] for the CVs obtained from the basic operations of superposition, splitting and departure, respectively, indicate, at least roughly, how a CV influences the rest of the network.

The variability of the departure process tends to be more seriously influenced by the variability of the service-time distribution at higher utilizations. This is illustrated by the approximation formula for the CV of a departure process from a single-server queue first proposed in [25] and further, discussed in [37], namely,

$$CV_d^2 = \rho^2 CV_s^2 + (1-\rho)^2 CV_a^2, \quad (1)$$

where CV_d, CV_s, and CV_a are the coefficients of variation, respectively, of an interdeparture time, a service time and an interarrival time, and $\rho = \lambda/\mu$ is the traffic intensity or utilization, defined as the ratio of the arrival rate λ and service rate μ.

It should be remembered, however, that the variability of a departure process or an internal arrival process is determined not only by the variability of each interval , which is usually characterized reasonably by the interval CV, but also by the dependence among successive intervals. Except for external arrival processes, these processes are typically not renewal processes. Positive dependence, e.g., positive correlations, among successive intervals (a tendency for arrivals or departures to occur in clusters) can be regarded as another aspect of increased variability. Similarly, negative dependence can be regarded as another aspect of decreased variability.

The influence of CV_a and CV_s on a single-server queue may be described quantitatively by the following approximation formula for the mean queue length (excluding the job in service):

$$E\,Q = \rho^2\left(CV_a^2 + CV_s^2\right)/2(1-\rho) \quad (2)$$

(see [33, equation (44)] and references therein).

3. Numerical evidence for closed networks

We now consider closed networks and discuss numerical evidence in support of principles (CS1)–(CS4), with particular attention to (CS3). These principles are illustrated in the exact global-balance solutions of small closed central-server networks in [1,24]. To describe these examples, let the demand at any queue equal the relative arrival rate or visit ratio (determined by solving the traffic-rate equations) multiplied by the mean service time. Since we are considering only single-server queues, the actual utilizations differ from the demands by a common multiplicative factor.

Balbo's central-server network is illustrated in Fig. 1. It consists of two queues with equal de-

Table 1 [a]
Performance measures for Balbo's network in Fig. 1 (from [1])

Parameters		CPU	DISK1	DISK2
Mean service time		0.028	0.04	0.28
Visit ratio		10.0	7.0	2.0
Demand		0.28	0.28	0.56
Model variant	**Congestion measure**	**CPU**	**DISK1**	**DISK2**
Product form	QL	0.89	0.89	4.23
	Utilization	0.486	0.486	0.972
	Throughput	17.35	12.15	3.47
CV = 0.6 at CPU	QL	0.83	0.86	4.32
	Utilization	0.488	0.488	0.977
	Throughput	17.44	12.21	3.49
CV = 2 at CPU	QL	1.02	0.96	4.03
	Utilization	0.477	0.477	0.953
	Throughput	17.03	11.92	3.41
CV = 5 at CPU	QL	1.25	0.96	3.79
	Utilization	0.446	0.445	0.889
	Throughput	15.88	11.11	3.18
CV = 10 at CPU	QL	1.38	0.85	3.78
	Utilization	0.416	0.416	0.832
	Throughput	14.87	10.41	2.97
CV = 0.6 at DISK1	QL	0.84	0.83	4.32
	Utilization	0.488	0.488	0.977
	Throughput	17.45	12.21	3.49
CV = 2 at DISK1	QL	0.95	1.00	3.95
	Utilization	0.477	0.477	0.954
	Throughput	17.03	11.92	3.41
CV = 10 at DISK1	QL	0.83	1.37	3.80
	Utilization	0.415	0.415	0.831
	Throughput	14.83	10.38	2.97
CV = 2 at DISK2	QL	1.00	1.00	4.00
	Utilization	0.465	0.465	0.931
	Throughput	16.62	11.64	3.32
CV = 5 at DISK2	QL	1.11	1.10	3.79
	Utilization	0.439	0.439	0.878
	Throughput	15.67	10.97	3.13
CV = 10 at DISK2	QL	1.13	1.13	3.73
	Utilization	0.431	0.431	0.863
	Throughput	15.41	10.79	3.08

[a] The network population is 6 in each case.

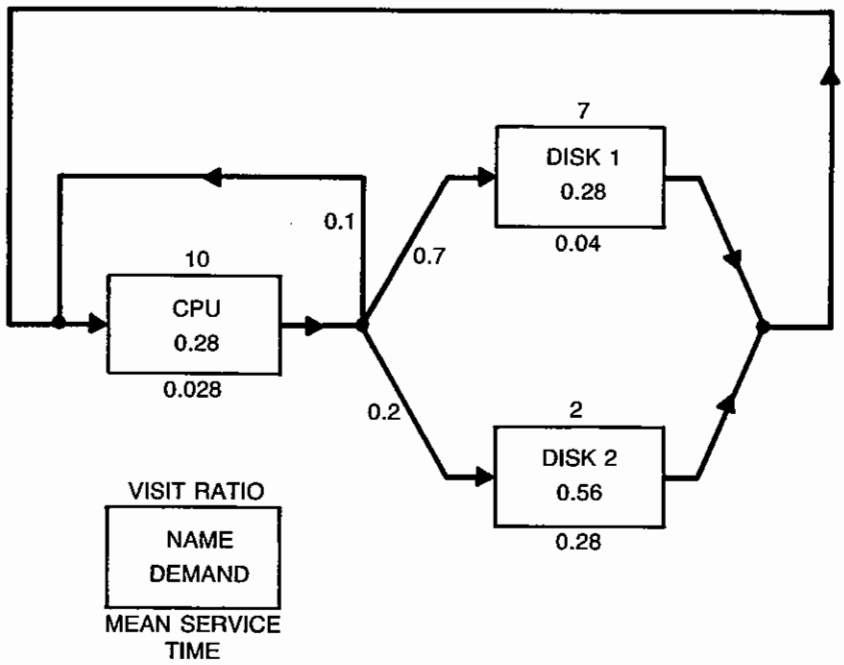

Fig. 1. Balbo's central-server model.

mands, CPU and DISK1, and a third queue, DISK2, whose demand is twice the demands of each of the other two queues. Clearly, DISK2 is the bottleneck device. The parameters of Balbo's central-server model and portions of the solution data are shown in Table 1. The network population (or multiprogramming level) is the same (6) in all cases. With this population, the system is saturated; e.g., the utilization of DISK2 when the network has all servers exponential is nearly one.

The reference case is the standard product-form model in which all service-time distributions are exponential. In all other cases, all the service-time distributions are again exponential except for the one distribution being modified. Servers with CV greater than one have been fitted with two-stage hyperexponential distributions chosen to have their stages balanced. The service-time density when the distribution has mean m and coefficient of variations $CV > 1$ is thus

$$f(t) = p\mu_1 e^{-\mu_1 t} + (1-p)\mu_2 e^{-\mu_2 t}, \quad t \geq 0, \qquad (3)$$

where

$$p = \tfrac{1}{2}\left(\sqrt{(CV^2-1)/(CV^2+1)}\right),$$

$$\mu_1 = 2p/m \quad \text{and} \quad \mu_2 = 2(1-p)/m.$$

Servers with CV less than one and greater than 0.5 have been fitted with hypoexponential distributions (convolutions of two exponential distributions, possibly having different means). Balbo solved this network exactly by representing it as a vector-valued Markov chain.

Table 1 illustrates principles (CS1)–(CS4). First, in support of (CS1), the mean queue lengths (denoted QL) change relatively little compared to what we would expect for related open models from approximation formulas (1) and (2). For example, the mean queue length would go up by a factor of 50.5 in an M/G/1 queue when the service-time CV is increased from 1 to 10.

Table 1 also shows the CPU mean queue length rising as its service-time CV increases with the service-time CVs of the other queues kept fixed at unity, while in contrast the mean queue length of DISK2 falls as its service-time CV is increased with other service time CVs fixed at unity. The utilizations at each server decrease whenever any service-time CV is increased.

Principles (CS1), (CS3), and (CS4) are also supported by the Ruggieri–Galeazzi data shown in Table 2. The network has the same topology as the Balbo model in Fig. 1, but different parameters. The mean queue length at DISK2 decreases when the DISK2 CV is increased as soon as the network population (MPL) exceeds 2. These data indicate that increasing the number of jobs makes

Table 2
The network in Fig. 1 with and without a hyperexponential service-time distribution at the bottleneck queue, taken from [24]

Parameters		CPU	DISK1	DISK2			
Mean service time		2.0	4.0	25.0			
Visit ratio		5.0	3.0	1.0			
Demand		10.0	12.0	25.0			
Model		All exponential			CV = 2.6 at DISK2		
Population	Measure	CPU	DISK1	DISK2	CPU	DISK1	DISK2
2	Mean queue length	0.37	0.46	1.17	0.37	0.46	1.18
	Utilization	0.305	0.367	0.764	0.289	0.346	0.721
	Throughput	0.153	0.092	0.031	0.144	0.087	0.029
3	Mean queue length	0.48	0.62	1.90	0.49	0.63	1.88
	Utilization	0.351	0.421	0.878	0.327	0.393	0.818
	Throughput	0.176	0.105	0.035	0.164	0.010	0.033
4	Mean queue length	0.55	0.73	2.72	0.59	0.79	2.62
	Utilization	0.375	0.450	0.937	0.350	0.420	0.876
	Throughput	0.187	0.112	0.037	0.175	0.105	0.035
5	Mean queue length	0.60	0.80	3.60	0.67	0.93	3.40
	Utilization	0.387	0.464	0.976	0.365	0.438	0.913
	Throughput	0.193	0.116	0.040	0.183	0.110	0.037
8	Mean queue length	0.66	0.90	6.45	0.83	1.23	5.93
	Utilization	0.398	0.478	0.996	0.387	0.466	0.965
	Throughput	0.199	0.120	0.040	0.194	0.116	0.039

(CS3) more prominent, but the phenomenon can occur if the system is only moderately loaded, as well as when it is saturated.

The apparent anomaly (CS3) was also investigated for the Balbo network in Fig. 1 using the SLAM II network-oriented simulation package [18]. As before all service-time distributions are exponential except for the designated one. The bottleneck phenomenon was also observed in the simulations. The simulations also showed a strong relationship between the anomalous mean queue lengths and the measured interarrival time CVs. The simulation results for the arrival and departure CVs appear in Table 3. No assessment of the statistical accuracy of the simulations is given, but the runs were quite long. Each network was simulated for 2000 units of simulated time on a CDC 6500, implying that there were at least 28 000 departures from the CPU. The statistics were cleared at time 200 to reduce startup bias. There were no replications with different seeds. However, the exact analytical results in Table 1 were used for validation.

From Table 3 we see that, as the CPU service-time CV is increased, the interarrival time CV at other devices in the system also increases. However, as suggested by (1), the same trend is much more marked when the service-time CV is increased at DISK2 instead. The difference is particularly great when the DISK2 CV is 10. This is also the instance in which CS3 is most prominent.

Table 3 is also significant because it dramatically demonstrates a striking difference between the open and closed models. The closed models is much more tightly coupled. The population constraint tends to induce strong negative dependence (e.g., correlations) among successive interarrival times at the queues. Indeed, if the successive interarrival times at each queue were nearly independent, then we could predict the mean queue lengths reasonably well using (2) together with the interarrival-time and service-time CVs. However, from Tables 1 and 3 we see that this method grossly overestimates the mean queue lengths. This shows that there must be significant negative dependence among successive interarrival times. For example, in the product-form case when the actual population is 6, the predicted expected number at the CPU using the open-model approximation (2) with $\rho = 0.486$ (the utilization from Table 1) and the simulation value $CV_a = 1.62$ (from Table 3) is 1.32 as opposed to the actual value 0.89 (Table 1). More dramatically, when CV = 10 at DISK2 and the actual population is 6, the predicted expected

Table 3
Arrival and departure CVs for Balbo's network in Fig. 1 and Table 1, obtained by simulation, from [2]

Parameters		CPU		DISK1		DISK2	
Service time		0.028		0.04		0.28	
Visit ratio		10.0		7.0		2.0	
Demand		0.28		0.28		0.56	
Variant	Population	ArrCV	DepCV	ArrCV	DepCV	ArrCV	DepCV
Product form	1	1.49	1.50	1.55	1.55	0.72	0.73
	6	1.62	1.63	1.67	1.67	0.98	0.97
CV = 0.6 at CPU	6	1.71	1.69	1.75	1.76	0.99	0.99
CV = 0.6 at DISK1	6	1.70	1.71	1.79	1.76	0.98	1.01
CV = 0.6 at DISK2	6	1.25	1.26	1.43	1.44	0.86	0.59
CV = 2 at CPU	1	1.55	1.55	1.60	1.60	0.74	0.74
	6	1.69	1.80	1.79	1.77	1.01	0.98
CV = 2 at DISK1	1	1.57	1.58	1.59	1.60	0.76	0.76
	6	1.77	1.75	1.74	1.86	1.04	0.99
CV = 2 at DISK2	1	2.71	2.71	2.42	2.42	1.24	1.24
	6	3.96	3.96	3.47	3.46	1.86	2.07
CV = 5 at CPU	1	1.98	1.98	1.92	1.92	0.93	0.93
	6	2.22	2.76	2.47	2.43	1.37	1.14
CV = 5 at DISK1	1	2.09	2.09	1.96	1.98	0.98	0.99
	6	2.42	2.42	2.21	2.79	1.37	1.25
CV = 5 at DISK2	1	4.73	4.73	4.08	4.09	2.15	2.17
	6	8.13	8.13	6.86	6.87	3.72	3.87
CV = 10 at CPU	1	2.76	2.80	2.55	2.54	1.26	1.27
	6	4.13	4.91	4.19	4.17	2.28	1.94
CV = 10 at DISK1	1	3.22	3.23	2.84	2.83	1.48	1.48
	6	4.15	4.15	3.58	4.22	2.13	1.97
CV = 10 at DISK2	1	13.24	13.24	11.09	11.09	5.94	5.98
	6	18.86	18.86	15.78	15.77	8.55	8.61

number at the CPU using (2) plus $\rho = 0.431$ (the utilization from Table 1) and the simulation value $CV_a = 18.86$ (from Table 3) is 58.7 as opposed to the actual value 1.13 (Table 1).

Two additional sets of experiments were run to explore principle (CS3) further. In the first set of experiments, a driver was written to run Marie's algorithm [15] repeatedly on a model whose mean service-time at a nonexponential server was gradually increased from zero. The program compared the approximate mean queue length at this server with the mean queue length the server would have had in a network with all servers exponential. As indicated in [1], Marie's approximation procedure is remarkably accurate. Moreover, principle (CS3) was always observed when the server's demand became sufficiently large. The same phenomenon was observed when the visit ratio was varied instead. The experiment was run with the service-time CV set at various levels. It was found that a single nonexponential server exhibited anomalous queue length at lower demands as the CV was raised. The results also showed that it is not necessary for the device to be the bottleneck in order to exhibit the anomaly ((CS3) is a rule of thumb, not a theorem) (see [2] for additional details).

The second experiment was performed on two-station cyclic networks. The networks had one exponential server and one server whose service-time CV was set to either one or five. The same hyperexponential service-time distributions in (3) were used. The steady-state equations were solved for networks with stations having balanced, slightly unbalanced, and very unbalanced demands, as shown in Table 4(a). All of the networks contained three circulating jobs. The balanced case shows a reduction in the queue length of the nonexponential server when its service-time CV is increased from 1 to 5, even though the utilization

Table 4 [a]
Results for a two-queue closed cyclic network with population 3 in support of principle (CS3)

(a) Global-balance solution of two-queue cyclic networks
queue 2 exponential

Network	CV_1	S_1	S_2	QL_1	QL_2	$Util_1$	$Util_2$	Throughput
Balanced	1.0	0.2	0.2	1.50	1.50	0.75	0.75	3.75
	5.0	0.2	0.2	1.46	1.54	0.65	0.65	3.24
Slightly	1.0	0.206	0.2	1.54	1.46	0.76	0.74	3.69
unbalanced	5.0	0.206	0.2	1.49	1.51	0.70	0.64	3.20
Very	1.0	0.2	0.1	2.27	0.73	0.93	0.47	4.67
unbalanced	5.0	0.2	0.1	2.15	0.85	0.86	0.43	4.32
	5.0	0.1	0.2	0.84	2.16	0.41	0.81	4.07

(b) Simulations of two-queue very unbalanced cyclic networks
queue 2 exponential

Cv_1	S_1	S_2	QL_1	QL_2	$Util_1$	$Util_2$	$CVarr_1$	$CVarr_2$	Throughput
1.0	0.2	0.1	2.26	0.74	0.94	0.47	0.99	0.98	4.72
5.0	0.2	0.1	2.16	0.84	0.86	0.43	4.37	4.39	4.27
5.0	0.1	0.2	0.84	2.16	0.41	0.81	2.09	2.18	4.04

[a] The mean service-time at node i is S_i and the mean queue length is QL_i, $i = 1, 2$.

of both servers is moderate (0.6480). The slightly unbalanced case is of interest because it shows that the longest queue length need not occur at the bottleneck device. This case provides an example in which one queue has both higher demand ($0.206 > 0.200$) and higher CV ($5 > 1$), but a lower mean queue length ($1.49 < 1.51$).

Another instance of the bottleneck anomaly is revealed by the very unbalanced case in Table 4(a). The analytic results for the very unbalanced case are corroborated by the simulation results in Table 4(b). Each network was run for 500 units of simulation time, after removing 100 units to reduce startup bias. Again, the exact analytical results were used for validation. Notice that the simulated CVs of the interarrival times for the network with one nonexponential server are considerably greater than those for the network with both servers exponential. The bottom row of Table 4(b) shows the results of a simulation in which the nonexponential server is *not* the bottleneck (the mean service times of the second row have been reversed); in this case, the mean queue length of the nonexponential server has increased when compared with that of the associated nonproduct form network. Also notice that, as suggested by (1), the CVs of the interarrival times are not as great as in the run shown in the second row of Table 4(b). This parallels the simulation results for Balbo's models; very large interarrival-time CVs were observed when a high service-time CV was placed at the bottleneck device.

4. Theoretical justification for principles (CS2) and (CS3)

There is a simple explanation for the service-time variability principles for closed networks when there is a large population (where large obviously depends on the network). As the population in the closed network grows, the bottleneck queue is almost always busy. (We assume that there is a single bottleneck queue, but the ideas also extend to several bottleneck queues.) In fact, it can be shown that the vector of queue-lengths in the subnetwork of the closed network without the bottleneck queue converges in distribution to the vector of queue lengths in the associated open network obtained by replacing the bottleneck queue with an external renewal arrival process. This and related limit theorems are discussed in [36]. The external renewal arrival process, of course, has the bottleneck service times as interarrival times. Hence, for large populations, we can regard the bottleneck service times as interarrival times to an open network. When the population is large enough, we can translate the service-time variability questions about closed networks into related arrival-process variability questions about

open networks. In particular, (CS2) and (CS3) can be viewed as consequences of (OS1) and (OA1), respectively, when the populations are large enough. This is rigorously justified asymptotically as the population grows, but the examples in Section 3 show that the effect is already present for relatively small populations.

In the case of a two-queue cyclic network, we can rigorously demonstrate that principles (CS2) and (CS3) for a closed network with a bottleneck node and a large population follow from principles (OS1) and (OA1). As the population grows, the closed network approaches a simple GI/G/1 queue with the bottleneck service times as interarrival times. For this special case, principles (CS2) and (CS3) follow from principles (OS1) and (OA1) via Theorems 2.1 and 2.2.

The discussion above indicates that principle (CS2) will eventually apply as the population grows, but it is not clear how large the population must be. From examples, the population evidently does not have to be very large. The population need not be so large when the relative utilization of the bottleneck node is large. This can be conveniently demonstrated in the special case of the two-queue cyclic network in which one of the service-time distributions is exponential. In this case, the closed model with population K is equivalent to the open $M/G/1/K-1$ model having a finite waiting room of size $K-1$ (see [20,13] and [12, p. 33]). We can see what happen in this case from the extensive tables in [12]. To illustrate, Tables 5 and 6 contains the mean queue lengths and utilizations for the server when $K=3$ and 11, respectively. In these tables, four service-time distributions with identical means are considered:

Table 5
The utilizations and mean queue lengths in an $M/G/1/K-1$ model with a finite waiting room of size $K-1=2$ for $G=D$ (Deterministic), E_2 (Erlang of order 2), M (exponential), and H_2 (hyperexponential).

Arrival rate	Utilization of the server				Mean queue length			
	D	E_2	M	H_2	D	E_2	M	H_2
0.50	0.486	0.476	0.467	0.451	0.18	0.23	0.27	0.31
0.75	0.685	0.656	0.634	0.606	0.42	0.48	0.51	0.55
1.00	0.824	0.781	0.750	0.716	0.70	0.73	0.75	0.77
1.40	0.938	0.893	0.859	0.826	1.08	1.05	1.05	1.04
2.00	0.988	0.960	0.933	0.909	1.42	1.36	1.33	1.32

Notes:
(1) The mean service time is always 1.
(2) The H_2 distribution has CV = 1.50 and balanced means.
(3) The exact values come from [12, Tables 5.1.2, 5.2.2, 5.3.2 and 5.4.8 in Section II.5].
(4) This model is equivalent to the closed two-queue cyclic network with population $K=3$ in which one service-time distribution is exponential. The mean service time of the exponential server is the reciprocal of the arrival rate.

Table 6
The utilizations and mean queue lengths in an $M/G/1/K-1$ model with a finite waiting room of size $K-1=10$ for $G=D$ (Deterministic), E_2 (Erlang of order 2), M (exponential), and H_2 (hyperexponential).

Arrival rate	Utilization of the server				Mean queue length			
	D	E_2	M	H_2	D	E_2	M	H_2
0.50	0.500	0.500	0.500	0.498	0.25	0.37	0.50	0.77
0.75	0.749	0.747	0.742	0.725	1.10	1.54	1.87	2.29
1.00	0.954	0.934	0.917	0.883	4.61	4.60	4.58	4.46
1.40	1.000	0.998	0.993	0.979	8.61	8.12	7.72	7.18
2.00	1.000	1.000	1.000	0.999	9.37	9.19	9.00	8.75

Notes:
(1) The mean service time is always 1.
(2) The H_2 distribution has CV = 1.50 and balanced means.
(3) The exact values from [12, Tables 5.1.6, 5.2.6, 5.3.6 and 5.4.12 in Section II.5] using the FCFS discipline.
(4) This model is equivalent to the closed two-queue cyclic network with population $K=11$ in which one service-time distribution is exponential. The mean service time of the exponential server is the reciprocal of the arrival rate.

D (deterministic), E_2 (Erlang of order 2), M (exponential) and H_2 (hyperexponential, a mixture of two exponentials with balanced means). The CVs are 0, 0.071, 1.0, and 1.50, respectively. These distributions are increasingly more variable in the $\leqslant_v$ ordering introduced in Section 2. Principles (CS1)–(CS4) are illustrated in these tables. Even when the population in 3, principle (CS3) applies. However, this is not the case when $K = 2$. The mean queue lengths when the arrival rate is 2.00 are then 0.53, 0.56, 0.57, and 0.58 for D, E_2, M, and H_2, respectively.

It is not necessary for the ratio of server utilizations to be extremely large. In fact, (CS3) holds for equal utilizations here (arrival rate = 1.00) when $K = 11$. We conjecture that there exist critical thresholds of populations and relative utilizations, such that principle (CS3) applies for all larger populations and all bottleneck nodes with higher relative utilizations.

5. Theoretical justification for principle (CS4)

Principle (CS4) is illustrated in all the examples so far. It can also be explained intuitively in the case of a bottleneck node with a large population. Then the throughput tends to be the bottleneck service rate. However, greater variability anywhere in the network increases the likelihood of the rare event that the bottleneck queue will be idle, thus decreasing the bottleneck utilization and the throughput. Now we give a theoretical justification of (CS4) in the special case of the two-queue cyclic closed network in which one service-time distribution is exponential. As in Section 2 we use the variability ordering $\leqslant_v$ based on the expectation of convex functions. As in Section 4, we use the equivalence with the M/G/1/$K-1$ systems.

Theorem 5.1. *In an* M/G/1/$K-1$ *queue for which the arrival rate is less than the service rate, the utilization is a nonincreasing function of the service-time distribution in the ordering* $\leqslant_v$.

Proof. By [6, Section 5.9], the utilization is $1/(\pi_0^* + a)$, where $a = \lambda/\mu$ is the offered load (arrival rate divided by the service rate),

$$\pi_0^* = p_0/(p_0 + \cdots + p_{K-1}),$$

and p_j is the equilibrium probability that there are j customers in the associated M/G/1 queue with unlimited waiting room. As a consequence, $p_0 = 1 - a$, independent of the service-time distribution. However, by Rolski and Stoyan [21], the equilibrium distribution of the M/G/1 queue increases stochastically when the service-time distribution gets more variable in the sense of $\leqslant_v$. Hence, $p_0 + \cdots + p_{K-1}$ decreases for any K, so that π_0^* increases and the utilization decreases. □

Remarks. (1) By the argument of Theorem 2.1, Theorem 5.1 can be recast in terms of the set of possible utilizations as a function of the service-time CV.

(2) The examples here indicate that server utilizations are not severely affected by service-time variability. This is typically the case, but dramatic changes in utilization are possible if the variability is allowed to increase without bound (see [36, Section XI]). It is possible to construct networks of single-server queues for which all server utilizations are arbitrarily close to 1 in the product-form Markovian model, but arbitrarily close to 0 in the non-Markovian model obtained simply by increasing the CVs of the service times while keeping the means fixed.

6. Approximation methods for closed networks

The results of the previous sections yield insights for evaluating the performance of approximate closed queueing network solution methods. Some of these methods will be examined below.

6.1. Reiser's MVA-based approximation

Reiser [19] has presented an approximate analysis of closed networks with nonexponential FCFS servers based on the mean-value-analysis (MVA) algorithm for product-form-networks. The algorithm is shown in Fig. 2.

The approximation reduces to the original (exact) MVA algorithm when all servers are exponential. The appearance of the squared service-time CV in the response-time formula is derived from a renewal argument based on the residual service time of the job in service at the instant a new job arrives. This is also the cause of its appearance in

Algorithm: REISER HEURISTIC

for i: = 1 to N do $\bar{n}_i(0) := 0.0$;
for k: = 1 to K do begin

$R_i := S_i(1 + \bar{n}_i(k-1) - U_i(k-1)) + U_i(k-1)S_i(1 + CV_i^2)/2$, $i = 1,2,\ldots,N$;

$\bar{n}_i(k) := kV_iR_i/\sum V_iR_i$, $i = 1,2,\ldots,N$;

end; (* for k*)

Notation:

K	network population (multiprogramming level)
N	number of queues
R_i	mean response time at queue i
S_i	mean service time at queue i
V_i	visit ratio or relative arrival rate
$\bar{n}_i(k)$	mean number of queue i given population k
$U_i(k)$	utilization of server i given population k

Fig. 2. Reiser's MVA-based heuristic.

the $P - K$ formula for the M/G/1 queue [10]. The approximation implicitly uses the Arrival Theorem [26], which states that the number of waiting jobs seen by an arriving job is equal to that seen by an outside observer of the network with one job removed. However, the Arrival Theorem does not hold unless the network has product form.

This approximation method may predict performance measures that are mutually inconsistent. To see this, consider the case when the network population is 2. Let $U_i(k)$ be the utilization and $\bar{n}_i(k)$ the expected number at queue i as a function of the network population k. Multiply the response-time formula given in Fig. 2 by the throughput on both sides and apply Little's law to obtain the following expression for the mean number of customers at node i:

$$\bar{n}_i(2) = U_i(2) + \tfrac{1}{2}U_i(2)U_i(1)\left(1 + CV_i^2\right), \qquad (4)$$

because $\bar{n}_i(1) = U_i(1)$, the utilization with one job in the system. The left-hand side of (4) is less than or equal to 2, the network population, while the right-hand side can be made arbitrarily large by letting $U_i(k)$ be the true utilization at node i with population k and increasing CV_i. To see that the right-hand side of (4) can indeed be made arbitrarily large, consider the special case of a symmetric two-node network in which both servers have the same CV. Since $U_1(1) + U_2(1) = 1$, $U_i(1) = \frac{1}{2}$ and $U_i(2) \geq \frac{1}{2}$. Hence, (4) implies that $\bar{n}_i(2) > 2$ for $CV > \sqrt{11}$. Of course, Reiser's algorithm does not actually make $\bar{n}_i(2) > 2$ when CV_i gets large; it forces $U_i(k)$ below its true value to keep $\bar{n}_i(2) < 2$.

The presence of the CV on the right-hand side of the response-time formula suggests (but does not demonstrate) that the queue length will always increase as the coefficient of variation increases, contradicting the numerical results of Section 3. There is also nothing in the approximation to suggest that the effect of the service-time CV on the queue length at one server is counteracted by the effect of interarrival-time variability on the queue length of other queues.

Sample outputs of Reiser's heuristic for Balbo's central-server model are shown in Table 7. Note that the predicted mean queue length at DISK2 goes up from 4.23 to 4.75 to 5.71 as the service time CV is increased from 1 to 2 to 10, contrary to principle (CS3). Also note that the approximate values are not close to the exact values. These observations do not conflict with Reiser's observation that the approximation produces good results

Table 7
Comparison of the Reiser, EPF, CHW, and Marie approximation methods for Balbo's central server model, as described in Table 1

CV = 1.0 at DISK2, product form			
Server	Queue length	Utilization	Throughput
CPU	0.89	0.486	17.35
DISK1	0.89	0.486	12.15
DISK2	4.23	0.972	3.47
CV = 2.0 at DISK2, Reiser's method			
Server	Queue length	Utilization	Throughput
CPU	0.63	0.397	14.17
DISK1	0.63	0.397	9.92
DISK2	4.75	0.793	2.83
CV = 2.0 at DISK2, EPF method			
Server	Queue length	Utilization	Throughput
CPU	0.99	0.467	16.68
DISK1	0.99	0.467	11.68
DISK2	4.02	0.934	3.34
CV = 2.0 at DISK2, CHW method			
Server	Queue length	Utilization	Throughput
CPU	0.87	0.479	17.10
DISK1	0.87	0.479	11.97
DISK2	4.22	0.947	3.38
CV = 2.0 at DISK2, Marie's method			
Server	Queue length	Utilization	Throughput
CPU	1.00	0.465	16.60
DISK1	1.00	0.465	11.62
DISK2	4.00	0.930	3.32
CV = 5.0 at DISK2, Reiser's method			
Server	Queue length	Utilization	Throughput
CPU	0.31	0.241	8.61
DISK1	0.31	0.241	6.03
DISK2	5.37	0.482	1.72
CV = 5.0 at DISK2, EPF method			
Server	Queue length	Utilization	Throughput
CPU	1.10	0.440	15.71
DISK1	1.10	0.440	11.00
DISK2	3.80	0.880	3.14
CV = 5.0 at DISK2, CHW method			
Server	Queue length	Utilization	Throughput
CPU	0.84	0.469	16.76
DISK1	0.84	0.469	11.73
DISK2	4.27	0.930	3.32
CV = 5.0 at DISK2, Marie's method			
Server	Queue length	Utilization	Throughput
CPU	1.11	0.438	15.66
DISK1	1.11	0.438	10.96
DISK2	3.78	0.877	3.13
CV = 10.0 at DISK2, Reiser's method			
Server	Queue length	Utilization	Throughput
CPU	0.15	0.123	4.38
DISK1	0.15	0.123	3.06
DISK2	5.71	0.245	0.88
CV = 10.0 at DISK2, EPF method			
Server	Queue length	Utilization	Throughput
CPU	1.13	0.432	15.42
DK1	1.13	0.432	10.79
DISK2	3.74	0.936	3.08
CV = 10.0 at DISK2, CHW method			
Server	Queue length	Utilization	Throughput
CPU	0.84	0.468	16.70
DISK1	0.84	0.468	11.69
DISK2	4.27	0.957	3.31
CV = 10.0 at DISK2, Marie's method			
Server	Queue length	Utilization	Throughput
CPU	1.13	0.431	15.40
DISK1	1.13	0.431	10.78
DISK2	3.73	0.931	3.08

for networks with one or more deterministic servers, since the nature of the variability in our example is quite different.

6.2. The extended-product-form (EPF) method

The Extended-Product-Form (EPF) method, due to Shum and Buzen [27,28] treats nonexponential servers by convolving the distributions of M/G/1/K queues into the normalizing constant vector G used in Buzen's algorithm [4]. Trial throughputs of the M/G/1/K queues are determined using a search whose objective is to minimize violations of the flow balance equations $X_i = V_i X_0$, $i = 1, 2, \ldots, K$. The method reduces to the original convolution algorithm when all servers are exponential. Balbo's thesis [1] indicates that EPF is fairly accurate in many cases. The portion of his data reproduced in Table 7 indicates that the EPF method predicts principle (CS3) correctly and is very accurate for this example. However, this should not be too surprising because the EPF technique is known to be exact for a two-queue cyclic network with only one nonexponential server [27,28], and Balbo's example in Fig. 1 is very close to such a network.

It is not difficult to see how the EPF can perform poorly. For example, consider a cyclic network with a bottleneck nonexponential server having CV $\gg 1$ and $n \geq 2$ other exponential servers with equal service rates. The EPF method necessarily makes the queue-length distributions the same at all n exponential servers, whereas in fact the mean queue lengths are typically much larger in the queues that come immediately after the bottleneck server. As the population grows, the closed network approaches the open network containing a series of exponential servers with a bursty external arrival process. The successive mean queue lengths at the exponential servers decrease dramatically if the traffic intensity is not too small. For $n \geq 2$ and large populations, it thus is apparent that the EPF method can perform poorly and presumably also violate (CS3).

6.3. The generalized-product-form (GPF) method

A generalization of the EPF method called the Generalized-Product-Form (GPF) method was developed by Tripathi [32]. It iteratively applies approximations for the interarrival and interdeparture time CV's as in (1) and [25] to compute arrival-process parameters for heuristic approximations for GI/G/1 queue-length distributions. These distributions are then convolved into the normalizing constant vector G. The vector G is then used to compute the network's performance measures in the usual way. Initial estimates for the throughputs are obtained using the Chandy/Herzog/Woo (CHW) decomposition method described in Section 6.4. We do not display results for this method, but the results in [32] were not especially good. Tripathi attributed errors in the output of this method to errors in the individual approximating queue-length distributions. It is also possible that errors may arise from the CV approximations for departure processes described earlier. However, we believe that the errors are primarily caused by ignoring the population constraint and the dependence it causes in the arrivals. Nevertheless, it does appear that the GPF method or a suitable refinement should usually be consistent with principle (CS3). Avoiding the Poisson arrival assumption should make it possible to improve on the GPF method.

6.4. The CHW device-complement method

The device-complement method of Chandy, Herzog and Woo (CHW) [5] decomposes a closed network into two parts, a (possibly) nonexponential device with index i and a single flow-equivalent server whose state-dependent service rate $\mu_i(k)$ is taken to be the throughput of the closed network with device i removed (the *complementary network*) and k circulating jobs. The complementary network is assumed to consist entirely of exponential servers so that it may be solved using the convolution algorithm. The global balance equations of the two-station network are then solved by an appropriate method. The service-time CV of the composite server is taken to be an average of the CV's of the servers it represents, weighted by their throughputs. After the decomposition and solution procedure have been performed for each station in turn, corrections to the service rates are made so as the minimize the violations of the flow-balance law and the queue-length constraints. The entire process is then repeated until the constraints are met to a set tolerance. A lucid description of the CHW method is given in [31].

The CHW decomposition method has been reported to predict throughputs fairly accurately [1], but not queue lengths. Errors may result from assuming exponential service at all stations of the flow-equivalent complementary server when computing its throughputs. Indeed, the data in [1] show that the CHW decomposition method does not predict (CS3). The cases CV = 2, 5, and 10 are displayed in Table 7. While the CHW method is not consistent with (CS3), it is much more accurate than Reiser's MVA approximation for this example.

6.5. Marie's device-complement method

Marie's method [15] is also an iterative device-complement method. Each device is treated as a queue with state-dependent Poisson arrivals and a Coxian server [7], denoted by $\lambda_i(k)/C/1$. Coxian servers consist of a series of exponential stages; a job leaves the server after each stage with some probability. Stages and exit probabilities may be combined to yield a Coxian server with any desired CV. The queue-length distribution is computed as described in [16] when the arrival rates

and Laplace transform $f_i(s)$ of the service-time distribution at queue i are known.

For closed networks with general servers, the arrival rates at server i are given by

$$\lambda_i(K) = 0$$

and

$$\lambda_i(k) = V_i \frac{G_i(K-k-1)}{G_i(K-k)}, \quad 0 \leqslant k \leqslant K, \tag{5}$$

where G_i denotes the convolution vector of the complementary network and K is the population or multiprogramming level. The queue-length distribution is used to fit an equivalent load-dependent server to device i when it is in the complementary networks of the other servers on the next iteration, since

$$\lambda_i(k-1)p_i(k-1) = \mu_i(k)p_i(k), \quad k = 1, 2, \ldots, K, \tag{6}$$

where $\lambda_i(k)$, $\mu_i(k)$, and $p_i(k)$ are respectively the arrival rate, service rate, and state probability for k customers at node i. This allows the complementary networks to be solved by the convolution method, while explicitly accounting for the distributional form of each server.

After the throughputs and queue lengths have been computed, they are checked to ensure that the flow-balance and queue-length constraints are satisfied. If they are not satisfied, corrections similar to those for the CHW method are applied to the service rates and the entire process is repeated. The overall scheme is depicted in Fig. 3. The corrections are described in [15,2] and will not be repeated here.

The data in [1,24] indicate that Marie's method is one of the most accurate approximations for closed single-class queueing networks with nonexponential servers, and this is corroborated here in Table 7. The accuracy is apparently achieved by explicitly accounting for the distributional form of the service time at all devices. This is done by (i) using the Laplace transform to compute the queue-length distribution at each queue, and (ii) fitting the service times with equivalent load-dependent service rates used in analyzing the complementary networks. Marie's algorithm also implicitly captures the effect of high or low interarrival-time variability by fitting instantaneous state-dependent arrival rates to each server. Because of these features, Marie's method seems to perform consistently well. However, more testing is needed to investigate its convergence properties and its performance over a wide range of service-time distributions.

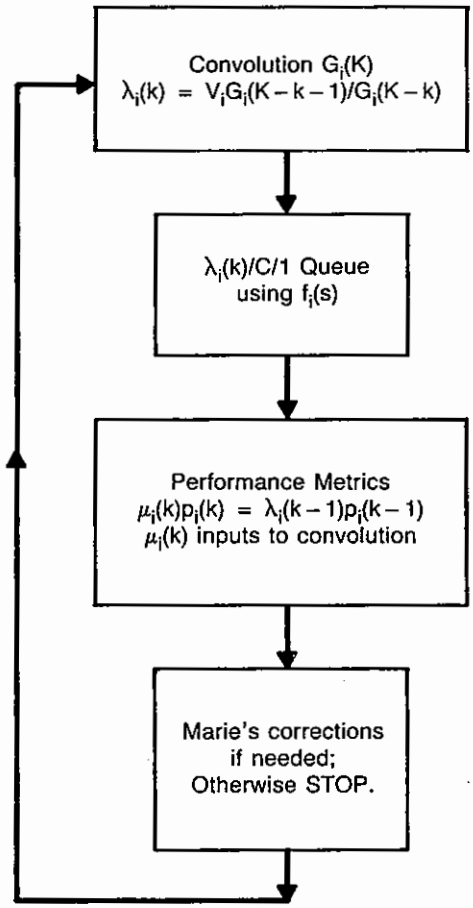

Fig. 3. Scheme for Marie's algorithm.

Acknowledgment

We thank Peter Denning, Cristina Ruggieri, Herb Schwetman, Moshe Segal and the referees for their valuable comments. Computing facilities for the first author were provided by the Purdue University Department of Computer Sciences and the Purdue University Computing Center.

References

[1] G. Balbo, Approximate solutions of queueing network models of computer systems, Ph.D. Dissertation, Dept. of Computer Sciences, Purdue Univ., 1979.

[2] A.B. Bondi, Incorporating open queueing models into closed queueing network algorithms, Ph.D. Dissertation, Dept. of Computer Science, Purdue Univ., 1984.
[3] W. Bux and U. Herzog, The phase concept: Approximation of measured data and performance analysis, in: K.M. Chandy and M. Reiser, eds., Computer Performance (North-Holland, Amsterdam/New York, 1977) 23–38.
[4] J.P. Buzen, Computational algorithms for closed queueing networks with exponential servers, Comm. ACM 16 (9) (1973) 527–531.
[5] K.M. Chandy, U. Herzog and L. Woo, Approximate analysis of general queueing networks, IBM J. Res. Develop. 19 (1) (1975) 43–49.
[6] R.B. Cooper, Introduction to Queueing Theory (North-Holland, Amsterdam, New York, 2nd ed., 1981).
[7] D.R. Cox, A use of complex probabilities in the theory of stochastic processes, Proc. Camb. Philos. Soc. 51 (1955) 313–319.
[8] W.J. Gordon and G.F. Newell, Closed queueing systems with exponential servers, Oper. Res. 15 (2) (1967) 254–65.
[9] J.R. Jackson, Jobshop-like queueing systems, Management Sci. 10 (1) (1963) 131–42.
[10] L. Kleinrock, Queueing Systems, Volume 1: Theory (Wiley, New York, 1975).
[11] J.G. Klincewicz and W. Whitt, On approximations for queues, II: Shape constraints, AT&T Bell Lab. Tech. J. 63 (1) (1984) 139–162.
[12] P. Kühn, Tables on delay systems, Rept., Institute of Switching and Data Technics, Univ. of Stuttgart, 1976.
[13] S.S. Lavenberg, The steady-state queueing time distribution for the M/G/1 finite capacity queue, Management Sci. 21 (5) (1975) 501–506.
[14] E.D. Lazowska, The use of percentiles in modelling CPU service time distributions, in: K.M. Chandy and M. Reiser, eds., Computer Performance (North-Holland, Amsterdam/New York, 1977) 53–66.
[15] R. Marie, Methodes iteratives de resolution de modeles mathematiques de systemes informatiques, R.A.I.R.O. Informatique/Computer Science 12 (2) (1978) 107–122.
[16] R. Marie, Calculating equilibrium probabilities for $\lambda(n)/C_k/1/N$ queues, Perform. Eval. Rev. 9 (1980) 117–125.
[17] S. Niu, On the comparison of waiting times in tandem queues, J. Appl. Probab. 18 (1981) 704–714.
[18] A.A.B. Pritsker and C.D. Pegden, Introduction to Simulation and SLAM (Halsted Press, New York, 1979).
[19] M. Reiser, A queueing network analysis of computer communication networks with window flow control, IEEE Trans. Comm. COM-27 (8) (1979) 1199–1209.
[20] M. Reiser and H. Kobayashi, The effects of service-time distributions on system performance, in: J.L. Rosenfeld, ed., Information Processing '74 (North-Holland, Amsterdam, 1974) 230–234
[21] T. Rolski and D. Stoyan, On the comparison of waiting times in GI/G/1 queues, Oper. Res. 24 (1) (1976) 197–200.
[22] S.M. Ross, Average delay in queues with non-stationary poisson arrivals, J. Appl. Probab. 15 (1978) 602–609.
[23] S.M. Ross, Stochastic Processes (Wiley, New York, 1983).
[24] C. Ruggieri and P. Galeazzi, Teoria della quasi completa decomponibilita' e applicazioni a modelli a reti di code, Tesi di Laurea (Baccalaureate Thesis), Universita degli studi di Torino, Torino, Italy, 1981.
[25] K.C. Sevcik, A.I. Levy, S.K. Tripathi and J.L. Zahorjan, Improving approximations of aggregated queueing network subsystems, in: K.M. Chandy and M. Reiser, eds., Computer Performance: Internat. Symp. on Computer Performance Modeling, Measurement, and Evaluation, Yorktown Heights, NY (North-Holland, Amsterdam, 1977) 1–22.
[26] K.C. Sevcik and I. Mitrani, The distribution of queueing network states at input and output instants, in: M. Arato, A. Butrimenko and E. Gelenbe, eds., Performance of Computer Systems (North-Holland, Amsterdam, 1979) 319–335.
[27] A.W. Shum, Queueing Models for Computer Systems with General Service Time Distributions (Garland, New York, 1980).
[28] A.W. Shum and J.P. Buzen, The EPF technique: A method for obtaining approximate solutions to closed queueing networks with general service times, Proc. 3rd Symp. on Measuring, Modeling, and Evaluating Computer Systems (North-Holland, Amsterdam, 1977) 201–220.
[29] D. Stoyan and H. Stoyan, Monotonicity properties of waiting times in the GI/G/1 Model, Zeit. Angew. Math. Mech. 49 (1969) 729–734 (in German).
[30] D. Stoyan, Comparison Methods for Queues and Other Stochastic Models (Wiley, New York, 1983) (English edition: edited by D.J. Daley, from: Qualitative Properties and Bounds for Stochastic Models (Akademie-Verlag, Berlin, 1977, in German).)
[31] S. Tolopka, Solution of general queueing networks using Norton's theorem, Tech. Rept. No. TR-314, Computer Sciences Dept., Purdue Univ., 1979.
[32] S.K. Tripathi, On approximate solution techniques for queueing network models of computer systems, Ph.D. Thesis, Tech. Rept. CSRG-106, C.S.R.G., Univ. of Toronto, 1981.
[33] W. Whitt, The queueing network analyzer, Bell System Tech. J. 62 (9) (1983) 2779–2815.
[34] W. Whitt, On approximations for queues, I: Extremal distributions, AT&T Bell Lab. Tech. J. 63 (1) (1984) 115–138.
[35] W. Whitt, Comparison conjectures about the M/G/s queue, Oper. Res. Lett. 2 (5) (1984) 203–209.
[36] W. Whitt, Open and closed models for networks of queues, AT&T Bell Lab. Tech. J. 63 (9) (1984) 1911–1979.
[37] W. Whitt, Approximations for departure processes and single-server queues in series, Nav. Res. Log. Quart. 31 (4) (1984) 499–521.
[38] R.W. Wolff, The effect of service time regularity on system performance, in: K.M. Chandy and M. Reiser, eds., Computer Performance (North-Holland, Amsterdam, 1977) 297–304.